Sketching as a Tool for Numerical Linear Algebra
All Lectures

David Woodruff
IBM Almaden

Massive data sets

Examples

- Internet traffic logs
- Financial data
- etc.

Algorithms

- Want nearly linear time or less
- Usually at the cost of a randomized approximation

Regression analysis

Regression

- Statistical method to study dependencies between variables in the presence of noise.

Regression analysis

Linear Regression

- Statistical method to study linear dependencies between variables in the presence of noise.

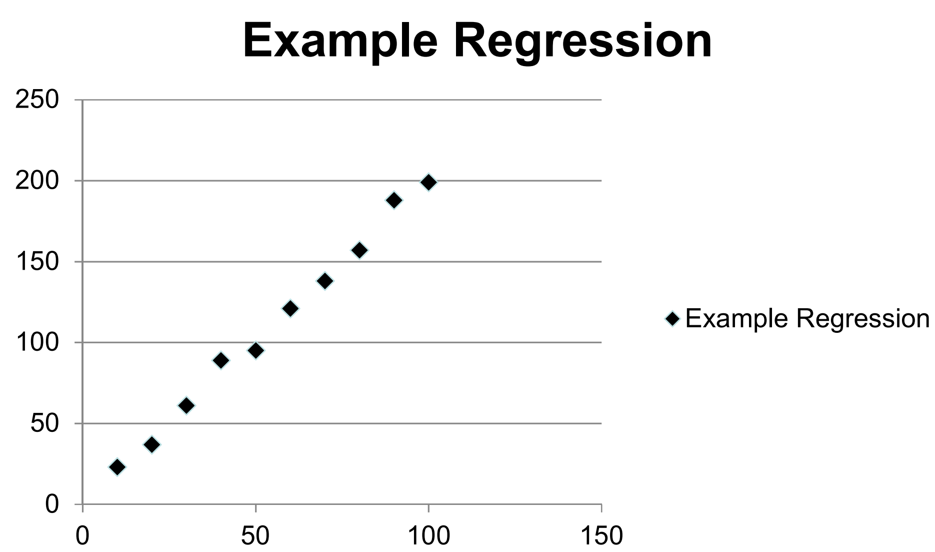
Regression analysis

Linear Regression

- Statistical method to study linear dependencies between variables in the presence of noise.

Example

- Ohm's law $V = R \cdot I$



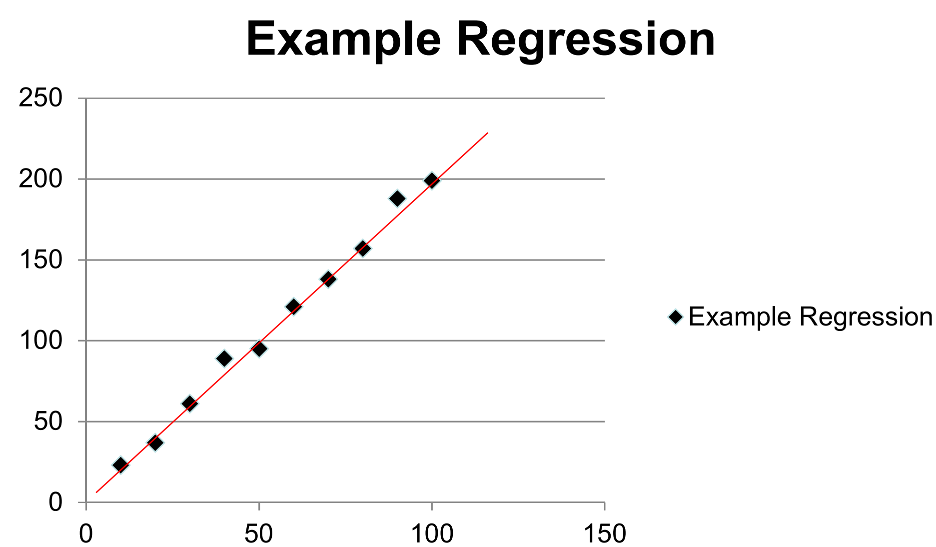
Regression analysis

Linear Regression

- Statistical method to study linear dependencies between variables in the presence of noise.

Example

- Ohm's law V = R · I
- Find linear function that best fits the data



Regression analysis

Linear Regression

- Statistical method to study linear dependencies between variables in the presence of noise.

Standard Setting

- One measured variable b
- A set of predictor variables $a_1, \ldots, a_d$
- Assumption:

$$b = x_0 + a_1 x_1 + \ldots + a_d x_d + \varepsilon$$

- ε is assumed to be noise and the x_i are model parameters we want to learn
- Can assume $x_0 = 0$
- Now consider n observations of b

Regression analysis

Matrix form

Input: $n \times d$-matrix A and a vector $b=(b_1,\ldots, b_n)$
n is the number of observations; d is the number of predictor variables

Output: x^* so that Ax^* and b are close

- Consider the over-constrained case, when $n \gg d$
- Can assume that A has full column rank

Regression analysis

Least Squares Method

- Find x* that minimizes $|Ax-b|_2^2 = \Sigma\ (b_i - \langle A_{i*}, x \rangle)^2$
- A_{i*} is i-th row of A
- Certain desirable statistical properties

Regression analysis

Geometry of regression

- We want to find an x that minimizes $|Ax-b|_2$
- The product Ax can be written as

$$A_{*1}x_1 + A_{*2}x_2 + ... + A_{*d}x_d$$

where A_{*i} is the i-th column of A

- This is a linear d-dimensional subspace
- The problem is equivalent to computing the point of the column space of A nearest to b in l_2-norm

Regression analysis

Solving least squares regression via the normal equations

- How to find the solution x to $\min_x |Ax-b|_2$?
- Equivalent problem: $\min_x |Ax-b|_2^2$
  - Write $b = Ax' + b'$, where b' orthogonal to columns of A
  - Cost is $|A(x-x')|_2^2 + |b'|_2^2$ by Pythagorean theorem
  - Optimal solution x if and only if $A^T(Ax-b) = A^T(Ax-Ax') = 0$
  - Normal Equation: $A^TAx = A^Tb$ for any optimal x
  - $x = (A^TA)^{-1} A^T b$
- If the columns of A are not linearly independent, the Moore-Penrose pseudoinverse gives a minimum norm solution x

Moore-Penrose Pseudoinverse

<u>Singular Value Decomposition (SVD)</u>

Any matrix A = U $\cdot$ Σ $\cdot$ V^T

- U has orthonormal columns
- Σ is diagonal with non-increasing non-negative entries down the diagonal
- V^T has orthonormal rows

- Pseudoinverse A^- = V $\Sigma^{-1}U^T$
  - Where Σ^{-1} is a diagonal matrix with i-th diagonal entry equal to $1/\Sigma_{ii}$ if $\Sigma_{ii} > 0$ and is 0 otherwise

- $\min_x |Ax-b|_2^2$ not unique when columns of A are linearly independent, but x = A^-b has minimum norm

Moore-Penrose Pseudoinverse

- Any optimal solution x has the form $\mathrm{A^{-}b} + \left(\mathrm{I} - \mathrm{V'V'^{T}}\right)\mathrm{z}$, where V' corresponds to the rows i of $\mathrm{V^{T}}$ for which $\Sigma_{i,i} > 0$

- Why?

- Because $\mathrm{A}\left(\mathrm{I} - \mathrm{V'V'^{T}}\right)\mathrm{z} = 0$, so $\mathrm{A^{-}b} + \left(\mathrm{I} - \mathrm{V'V'^{T}}\right)\mathrm{z}$ is a solution. This is a d-rank(A) dimensional affine space so it spans all optimal solutions

- Since $\mathrm{A^{-}b}$ is in column span of V', by Pythagorean theorem, $|\mathrm{A^{-}b} + \left(\mathrm{I} - \mathrm{V'V'^{T}}\right)\mathrm{z}|_2^2 = |\mathrm{A^{-}b}|_2^2 + |(\mathrm{I} - \mathrm{V'V'^{T}})\mathrm{z}|_2^2 \geq |\mathrm{A^{-}b}|_2^2$

Time Complexity

Solving least squares regression via the normal equations

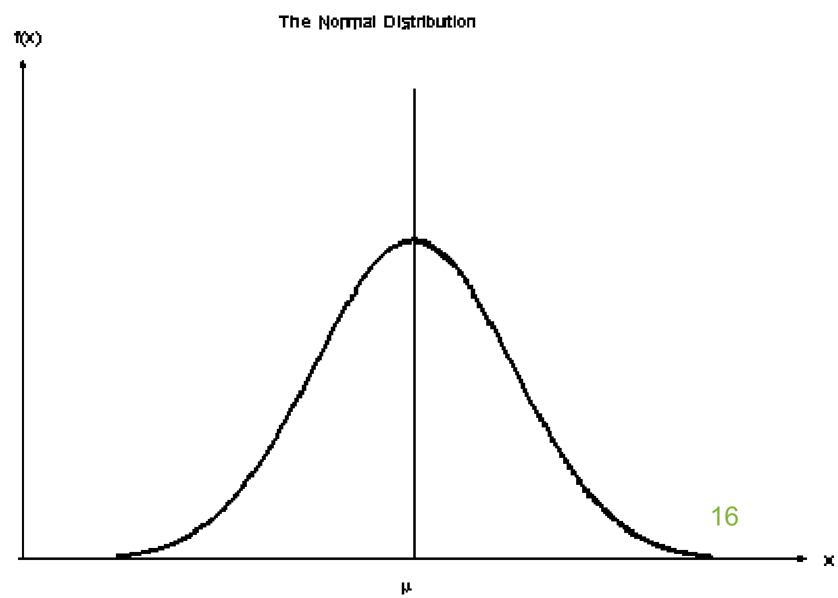
- Need to compute $x = A^{-}b$
- Naively this takes nd^2 time
- Can do $\mathrm{nd}^{1.376}$ using fast matrix multiplication
- But we want much better running time!

Sketching to solve least squares regression

- How to find an approximate solution x to $\min_x |Ax-b|_2$?
- Goal: output x' for which $|Ax'-b|_2 \leq (1+\varepsilon) \min_x |Ax-b|_2$ with high probability
- Draw S from a k x n random family of matrices, for a value k << n
- Compute S*A and S*b
- Output the solution x' to $\min_{x'} |(SA)x-(Sb)|_2$
  - $x' = (SA)^{-}Sb$

How to choose the right sketching matrix S?

- Recall: output the solution x‘ to $\min_{x'} |(SA)x-(Sb)|_2$
- Lots of matrices work
- S is d/ε^2 x n matrix of i.i.d. Normal random variables
- To see why this works, we introduce the notion of a subspace embedding



Subspace Embeddings

- Let $k = O(d/\varepsilon^2)$
- Let S be a k x n matrix of i.i.d. normal $N(0,1/k)$ random variables
- For any fixed d-dimensional subspace, i.e., the column space of an n x d matrix A
  - W.h.p., for all x in R^d, $|SAx|_2 = (1\pm\varepsilon)|Ax|_2$
- Entire column space of A is preserved

Why is this true?

Subspace Embeddings – A Proof

- Want to show $|SAx|_2 = (1\pm\varepsilon)|Ax|_2$ for all x

- Can assume columns of A are orthonormal (since we prove this for all x)

- Claim: SA is a k x d matrix of i.i.d. N(0,1/k) random variables

  – First property: for two independent random variables X and Y, with X drawn from $N(0,a^2)$ and Y drawn from $N(0,b^2)$, we have X+Y is drawn from $N(0, a^2 + b^2)$

X+Y is drawn from N(0, $a^2 + b^2$)

- Probability density function f_z of Z = X+Y is convolution of probability density functions f_X and f_Y

- $f_Z(z) = \int f_Y(z-x) f_X(x)\, dx$

- $f_x(x) = \frac{1}{a(2\pi)^{.5}} e^{-x^2/2a^2}$, $f_y(y) = \frac{1}{b(2\pi)^{.5}} e^{-x^2/2b^2}$

- $f_Z(z) = \int \frac{1}{a(2\pi)^{.5}} e^{-(z-x)^2/2a^2} \frac{1}{b(2\pi)^{.5}} e^{-x^2/2b^2} dx$

$$= \frac{1}{(2\pi)^{.5}(a^2+b^2)^{.5}} e^{-z^2/2(a^2+b^2)} \int \frac{\left(a^2+b^2\right)^{.5}}{(2\pi)^{.5}ab} e^{-\frac{\left(x-\frac{b^2 z}{a^2+b^2}\right)^2}{2\left(\frac{(ab)^2}{a^2+b^2}\right)}} \text{dx}$$

X+Y is drawn from N(0, $a^2 + b^2$)

Calculation: $e^{-\frac{(z-x)^2}{2a^2} - \frac{x^2}{2b^2}} = e^{-\frac{z^2}{2(a^2+b^2)} - \frac{\left(x-\frac{b^2 z}{a^2+b^2}\right)^2}{2\left(\frac{(ab)^2}{a^2+b^2}\right)}}$

Density of Gaussian distribution: $\int \frac{(a^2+b^2)^{.5}}{(2\pi)^{.5}ab} e^{-\frac{\left(x-\frac{b^2 z}{a^2+b^2}\right)^2}{2\left(\frac{(ab)^2}{a^2+b^2}\right)}}$ dx = 1

Rotational Invariance

- Second property: if u, v are vectors with <u, v> = 0, then <g,u> and <g,v> are independent, where g is a vector of i.i.d. N(0,1/k) random variables
- Why?
- If g is an n-dimensional vector of i.i.d. N(0,1) random variables, and R is a fixed matrix, then the probability density function of Rg is

$$f \quad (x) = \frac{1}{\det(\mathrm{RR^T})(2\pi)^{d/2}} e^{-\frac{x^T\left(\mathrm{RR^T}\right)^{-1} x}{2}}$$

  – $\mathrm{RR^T}$ is the covariance matrix
  – For a rotation matrix R, the distribution of Rg and of g are the same

Orthogonal Implies Independent

- Want to show: if u, v are vectors with <u, v> = 0, then <g,u> and <g,v> are independent, where g is a vector of i.i.d. N(0,1/k) random variables

- Choose a rotation R which sends u to αe_1, and sends v to βe_2

- $< g, u > = < gR, R^T u > = < h, \alpha e_1 > = \alpha h_1$
- $< g, v > = < gR, R^T v > = < h, \beta e_2 > = \beta h_2$

  where h is a vector of i.i.d. N(0, 1/k) random variables

- Then h_1 and h_2 are independent by definition

Where were we?

- Claim: SA is a k x d matrix of i.i.d. N(0,1/k) random variables

- Proof: The rows of SA are independent

  – Each row is: $< \mathrm{g}, \mathrm{A}_1 >, < \mathrm{g}, \mathrm{A}_2 >, \ldots, < \mathrm{g}, \mathrm{A}_\mathrm{d} >$

  – First property implies the entries in each row are N(0,1/k) since the columns A_i have unit norm

  – Since the columns A_i are orthonormal, the entries in a row are independent by our second property

Back to Subspace Embeddings

- Want to show $|SAx|_2 = (1 \pm \varepsilon)|Ax|_2$ for all x
- Can assume columns of A are orthonormal
- Can also assume x is a unit vector
- SA is a k x d matrix of i.i.d. N(0,1/k) random variables

- Consider any fixed unit vector $x \in R^d$
- $|SAx|_2^2 = \sum_{i \in [k]} < g_i, x >^2$, where g_i is i-th row of SA
- Each $< g_i, x >^2$ is distributed as $N\left(0, \frac{1}{k}\right)^2$
- E[$< g_i, x >^2$] = 1/k, and so E[$|SAx|_2^2$] = 1

How concentrated is $|SAx|_2^2$ about its expectation?

Johnson-Lindenstrauss Theorem

- Suppose $h_1, \ldots, h_k$ are i.i.d. N(0,1) random variables
- Then G = $\sum_i h_i^2$ is a χ^2-random variable
- Apply known tail bounds to G:
  - (Upper) $\Pr[G \geq k + 2(kx)^{.5} + 2x] \leq e^{-x}$
  - (Lower) $\Pr[G \leq k - 2(kx)^{.5}] \leq e^{-x}$
- If $x = \frac{\epsilon^2 k}{16}$, then $\Pr[G \in k(1 \pm \epsilon)] \geq 1 - 2e^{-\epsilon^2 k/16}$
- If $k = \Theta(\epsilon^{-2} \log(\frac{1}{\delta}))$, this probability is 1-δ

- $\Pr[|SAx|_2^2 \in (1 \pm \epsilon)] \geq 1 - 2^{-\Theta(d)}$

This only holds for a fixed x, how to argue for all x?

Net for Sphere

- Consider the sphere $S^{\mathrm{d}-1}$
- Subset N is a γ-net if for all $\mathrm{x} \in S^{\mathrm{d}-1}$, there is a $\mathrm{y} \in \mathrm{N}$, such that $|\mathrm{x}-\mathrm{y}|_2 \leq \gamma$
- Greedy construction of N
  - While there is a point $\mathrm{x} \in S^{\mathrm{d}-1}$ of distance larger than γ from every point in N, include x in N
- The sphere of radius γ/2 around every point in N is contained in the sphere of radius 1+ γ/2 around 0^{d}
- Further, all such spheres are disjoint
- Ratio of volume of d-dimensional sphere of radius 1+ $\gamma/2$ to dimensional sphere of radius γ is $(1+\gamma/2)^{\mathrm{d}}/(\gamma/2)^{\mathrm{d}}$, so $|\mathrm{N}| \leq (1+\gamma/2)^{\mathrm{d}}/(\gamma/2)^{\mathrm{d}}$

Net for Subspace

- Let M = {Ax | x in N}, so $|\mathrm{M}| \leq (1+\gamma/2)^{\mathrm{d}}/(\gamma/2)^{\mathrm{d}}$

- Claim: For every x in $S^{\mathrm{d}-1}$, there is a y in M for which $|\mathrm{Ax}-\mathrm{y}|_2 \leq \gamma$

- Proof: Let x' in $S^{\mathrm{d}-1}$ be such that $|\mathrm{x}-\mathrm{x}'|_2 \leq \gamma$

  Then $|\mathrm{Ax}-\mathrm{Ax}'|_2 = |\mathrm{x}-\mathrm{x}'|_2 \leq \gamma$, using that the columns of A are orthonormal. Set y = Ax'

Net Argument

- For a fixed unit x, $\Pr[|SAx|_2^2 \in (1 \pm \epsilon)] \geq 1 - 2^{-\Theta(d)}$
- For a fixed pair of unit x, x', $|SAx|_2^2$, $|SAx'|_2^2$, $|SA(x - x')|_2^2$ are all $1 \pm \epsilon$ with probability $1 - 2^{-\Theta(d)}$
- $|SA(x - x')|_2^2 = |SAx|_2^2 + |SAx'|_2^2 - 2 < SAx, SAx' >$
- $|A(x - x')|_2^2 = |Ax|_2^2 + |Ax'|_2^2 - 2 < Ax, Ax' >$
  - So $\Pr[< Ax, Ax' > = < SAx, SAx' > \pm O(\epsilon)] = 1 - 2^{-\Theta(d)}$
- Choose a ½-net M = {Ax | x in N} of size 5^d
- By a union bound, for all pairs y, y' in M,
$$< y, y' > = < Sy, Sy' > \pm O(\epsilon)$$
- Condition on this event
- By linearity, if this holds for y, y' in M, for $\alpha y, \beta y'$ we have
$$< \alpha y, \beta y' > = \alpha\beta < Sy, Sy' > \pm O(\epsilon\, \alpha\beta)$$

Finishing the Net Argument

- Let y = Ax for an arbitrary $x \in S^{d-1}$
- Let $y_1 \in M$ be such that $|y - y_1|_2 \leq \gamma$
- Let α be such that $|\alpha(y - y_1)|_2 = 1$
  - $\alpha \geq 1/\gamma$ (could be infinite)
- Let $y_2' \in M$ be such that $|\alpha(y - y_1) - y_2'|_2 \leq \gamma$
- Then $\left|y - y_1 - \frac{y_2'}{\alpha}\right|_2 \leq \frac{\gamma}{\alpha} \leq \gamma^2$
- Set $y_2 = \frac{y_2'}{\alpha}$. Repeat, obtaining $y_1, y_2, y_3, \ldots$ such that for all integers i,
$$|y - y_1 - y_2 - \ \ldots - y_i|_2 \leq \gamma^i$$
- Implies $|y_i|_2 \leq \gamma^{i-1} + \gamma^i \leq 2\gamma^{i-1}$

Finishing the Net Argument

- Have $y_1, y_2, y_3, \ldots$ such that $y = \sum_i y_i$ and $|y_i|_2 \leq 2\gamma^{i-1}$

- $|Sy|_2^2 = |S \sum_i y_i|_2^2$

  $= \sum_i |Sy_i|_2^2 + 2 \sum_{i,j} < Sy_i, Sy_j >$

  $= \sum_i |y_i|_2^2 + 2 \ \sum_{i,j} \ < y_i, y_j > \pm O(\epsilon) \sum_{i,j} |y_i|_2 \left|y_j\right|_2$

  $= | \sum_i y_i |_2^2 \ \pm O(\epsilon)$

  $= |y|_2^2 \pm O(\epsilon)$

  $= 1 \pm O(\epsilon)$

- Since this held for an arbitrary y = Ax for unit x, by linearity it follows that for all x, $|SAx|_2 = (1 \pm \varepsilon)|Ax|_2$

Back to Regression

- We showed that S is a subspace embedding, that is, simultaneously for all x,

$$|SAx|_2 = (1 \pm \varepsilon)|Ax|_2$$

What does this have to do with regression?

Subspace Embeddings for Regression

- Want x so that $|Ax-b|_2 \leq (1+\varepsilon) \min_y |Ay-b|_2$
- Consider subspace L spanned by columns of A together with b
- Then for all y in L, $|Sy|_2 = (1 \pm \varepsilon) |y|_2$
- Hence, $|S(Ax-b)|_2 = (1 \pm \varepsilon) |Ax-b|_2$ for all x
- Solve $\text{argmin}_y |(SA)y - (Sb)|_2$
- Given SA, Sb, can solve in poly(d/ε) time

Only problem is computing SA takes $O(nd^2)$ time

How to choose the right sketching matrix S? [S]

- S is a Subsampled Randomized Hadamard Transform
  - S = P*H*D
  - D is a diagonal matrix with +1, -1 on diagonals
  - H is the Hadamard transform
  - P just chooses a random (small) subset of rows of H*D
  - S*A can be computed in O(nd log n) time

Why does it work?

Why does this work?

- We can again assume columns of A are orthonormal
- It suffices to show $|\mathrm{SAx}|_2^2 = |\mathrm{PHDAx}|_2^2 = 1 \pm \epsilon$ for all x
- HD is a rotation matrix, so $|\mathrm{HDAx}|_2^2 = |\mathrm{Ax}|_2^2 = 1$ for any x
  - Notation: let y = Ax
- Flattening Lemma: For any fixed y,

$$\Pr\left[|\mathrm{HDy}|_\infty \geq C \frac{\log^{.5} \mathrm{nd}/\delta}{\mathrm{n}^{.5}}\right] \leq \frac{\delta}{2\mathrm{d}}$$

Proving the Flattening Lemma

- Flattening Lemma: Pr $[\|HDy\|_\infty \geq C \; \frac{\log^{.5} nd/\delta}{n^{.5}}] \leq \frac{\delta}{2d}$
- Let C > 0 be a constant. We will show for a fixed i in [n],

$$\Pr [|(HDy)_i| \;\; \geq C \; \frac{\log^{.5} nd/\delta}{n^{.5}}] \leq \frac{\delta}{2nd}$$

- If we show this, we can apply a union bound over all i
- $|(HDy)_i| = \sum_j H_{i,j} D_{j,j} y_j$
- (Azuma-Hoeffding) $\Pr\left[|\sum_j Z_j| > t\right] \leq 2e^{-\left(\frac{t^2}{2\sum_j \beta_j^2}\right)}$, where $|Z_j| \leq \beta_j$ with probability 1
  - $Z_j = H_{i,j} D_{j,j} y_j$ has 0 mean
  - $|Z_j| \leq \frac{|y_j|}{n^{.5}} = \beta_j$ with probability 1
  - $\sum_j \beta_j^2 = \frac{1}{n}$
- $\Pr\left[|\sum_j Z_j| > \frac{C \log^{.5}\left(\frac{nd}{\delta}\right)}{n^{.5}}\right] \leq 2e^{-\frac{C^2 \log\left(\frac{nd}{\delta}\right)}{2}} \leq \frac{\delta}{2nd}$

Consequence of the Flattening Lemma

- Recall columns of A are orthonormal
- HDA has orthonormal columns
- Flattening Lemma implies $|\mathrm{HDAe_i}|_\infty \leq C \ \frac{\log^{.5} \mathrm{nd}/\delta}{\mathrm{n}^{.5}}$ with probability $1 - \frac{\delta}{2\mathrm{d}}$ for a fixed $\mathrm{i} \in [\mathrm{d}]$
- With probability $1 - \frac{\delta}{2}$, $\left|\mathrm{e_j HDAe_i}\right| \ \leq C \ \frac{\log^{.5} \mathrm{nd}/\delta}{\mathrm{n}^{.5}}$ for all i,j
- Given this, $\left|\mathrm{e_j HDA}\right|_2 \leq C \ \frac{\mathrm{d}^{.5}\log^{.5} \mathrm{nd}/\delta}{\mathrm{n}^{.5}}$ for all j

(Can be optimized further)

Matrix Chernoff Bound

- Let $X_1, \ldots, X_s$ be independent copies of a symmetric random matrix $X \in R^{dxd}$ with $E[X] = 0$, $|X|_2 \leq \gamma$, and $\left|E[X^T X]\right|_2 \leq \sigma^2$. Let $W = \frac{1}{s}\sum_{i \in [s]} X_i$. For any $\epsilon > 0$,

$$\Pr[|W|_2 > \epsilon] \leq 2d \cdot e^{-s\epsilon^2/(\sigma^2 + \frac{\gamma\epsilon}{3})}$$

(here $|W|_2 = \sup |Wx|_2 / |x|_2$)

- Let V = HDA, and recall V has orthonormal columns
- Suppose P in the S = PHD definition samples uniformly with replacement. If row i is sampled in the j-th sample, then $P_{j,i} = n$, and is 0 otherwise
- Let Y_i be the i-th sampled row of V = HDA
- Let $X_i = I_d - n \cdot Y_i^T Y_i$
  - $E[X_i] = I_d - n \cdot \sum_j \left(\frac{1}{n}\right) V_j^T V_j = I_d - V^T V = 0^d$
  - $|X_i|_2 \leq |I_d|_2 + n \cdot \max \left|e_j HDA\right|_2^2 = 1 + n \cdot C^2 \log\left(\frac{nd}{\delta}\right) \cdot \frac{d}{n} = \Theta(d \log\left(\frac{nd}{\delta}\right))$

Matrix Chernoff Bound

- Recall: let Y_i be the i-th sampled row of V = HDA
- Let $X_i = I_d - n \cdot Y_i^T Y_i$
- $E[X^T X + I_d] = I_d + \ I_d - 2n\, E\left[Y_i^T Y_i\right] + n^2 E[Y_i^T Y_i Y_i^T Y_i]$

$$= 2I_d - 2I_d + n^2 \sum_i \left(\frac{1}{n}\right) \cdot v_i^T v_i v_i^T v_i = n \sum_i v_i^T v_i \cdot |v_i|_2^2$$

- Define $Z = n\ \sum_i v_i^T v_i\ C^2 \log\left(\frac{nd}{\delta}\right) \cdot \frac{d}{n} = C^2 d \log\left(\frac{nd}{\delta}\right)\ I_d$
- Note that $E[X^T X + I_d]$ and Z are real symmetric, with non-negative eigenvalues
- Claim: for all vectors y, we have: $y^T E[X^T X + I_d] y\ \le y^T Z y$
- Proof: $y^T E[X^T X + I_d]\, y = n\ \sum_i y^T v_i^T v_i y\, |v_i|_2^2 = n\ \sum_i < v_i, y >^2\ |v_i|_2^2$ and

$$y^T Z y = n \sum_i y^T v_i^T v_i y\, C^2 \log\left(\frac{nd}{\delta}\right) \cdot \frac{d}{n} = d \sum_i < v_i, y >^2\ C^2 \log\left(\frac{nd}{\delta}\right)$$

- Hence, $\left|E[X^T X]\right|_2 \le \left|E[X^T X] + I_d\right|_2 + |I_d|_2 = |E[X^T X + I_d]|_2 + 1$

$$\le |Z|_2 + 1 \le C^2 d \log\left(\frac{nd}{\delta}\right) + 1$$

- Hence, $\left|E[X^T X]\right|_2 = O\left(d \log\left(\frac{nd}{\delta}\right)\right)$

Matrix Chernoff Bound

- Hence, $\left|E[X^T X]\right|_2 = O\left(d \log\left(\frac{nd}{\delta}\right)\right)$

- Recall: (Matrix Chernoff) Let $X_1, \dots, X_s$ be independent copies of a symmetric random matrix $X \in R^{dxd}$ with $E[X] = 0$, $|X|_2 \le \gamma$, and $\left|E[X^T X]\right|_2 \le \sigma^2$. Let $W = \frac{1}{s}\sum_{i \in [s]} X_i$. For any $\epsilon > 0$, $\Pr[|W|_2 > \epsilon] \le 2d \cdot e^{-s\epsilon^2/(\sigma^2 + \frac{\gamma\epsilon}{3})}$

$$\Pr\left[\left|I_d - (PHDA)^T (PHDA) \quad \right|_2 > \epsilon\right] \le 2d \cdot e^{-s\,\epsilon^2/(\Theta(d \log\left(\frac{nd}{\delta}\right))}$$

- Set $s = d \log\left(\frac{nd}{\delta}\right) \frac{\log\left(\frac{d}{\delta}\right)}{\epsilon^2}$, to make this probability less than $\frac{\delta}{2}$

SRHT Wrapup

- Have shown $\left|I_d - (\mathrm{PHDA})^T(\mathrm{PHDA})\right|_2 < \epsilon$ using Matrix Chernoff Bound and with $s = \mathrm{d}\log\left(\frac{\mathrm{nd}}{\delta}\right)\frac{\log\left(\frac{\mathrm{d}}{\delta}\right)}{\epsilon^2}$ samples

- Implies for every unit vector x,

  $$\left|1 - |\mathrm{PHDAx}|_2^2\right| = \left|\mathrm{x}^T\mathrm{x} - \mathrm{x}^T(\mathrm{PHDA})^T(\mathrm{PHDA})\mathrm{x}\right| < \epsilon\,,$$

  so $|\mathrm{PHDAx}|_2^2 \in 1 \pm \epsilon$ for all unit vectors x

- Considering the column span of A adjoined with b, we can again solve the regression problem

- The time for regression is now only O(nd log n) + $\mathrm{poly}\left(\frac{d\log(n)}{\epsilon}\right)$. Nearly optimal in matrix dimensions (n >> d)

Faster Subspace Embeddings S [CW,MM,NN]

- CountSketch matrix
- Define k x n matrix S, for k = $O(d^2/\varepsilon^2)$
- S is really sparse: single randomly chosen non-zero entry per column

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Can compute $S \cdot A$ in nnz(A) time!

- nnz(A) is number of non-zero entries of A

Simple Proof [Nguyen]

- Can assume columns of A are orthonormal

- Suffices to show $|SAx|_2 = 1 \pm \varepsilon$ for all unit x
  - For regression, apply S to [A, b]

- SA is a $2d^2/\varepsilon^2$ x d matrix

- Suffices to show $|A^T S^T SA - I|_2 \leq |A^T S^T SA - I|_F \leq \varepsilon$

- Matrix product result shown below:

  $\Pr[|CS^TSD - CD|_F^2 \leq [6/(\delta(\text{\# rows of S}))] * |C|_F^2 |D|_F^2] \geq 1 - \delta$

- Set $C = A^T$ and $D = A$.
- Then $|A|^2_F = d$ and (# rows of S) $= 6\, d^2/(\delta\varepsilon^2)$

Matrix Product Result [Kane, Nelson]

- Show: $\Pr[|CS^TSD - CD|_F^2 \leq [6/(\delta(\#\text{ rows of } S))] * |C|_F^2 |D|_F^2] \geq 1 - \delta$

- (JL Property) A distribution on matrices $S \in R^{k x\, n}$ has the (ϵ, δ, ℓ)-JL moment property if for all $x \in R^n$ with $|x|_2 = 1$,
$$E_S \left| |Sx|_2^2 - 1 \right|^{\ell} \leq \epsilon^{\ell} \cdot \delta$$

- (From vectors to matrices) For $\epsilon, \delta \in \left(0, \frac{1}{2}\right)$, let D be a distribution on matrices S with k rows and n columns that satisfies the (ϵ, δ, ℓ)-JL moment property for some $\ell \geq 2$. Then for A, B matrices with n rows,

$$\Pr_S \left[\left| A^T S^T SB - A^T B \right|_F \geq 3\, \epsilon |A|_F |B|_F \right] \leq \delta$$

From Vectors to Matrices

- (From vectors to matrices) For $\epsilon, \delta \in \left(0, \frac{1}{2}\right)$, let D be a distribution on matrices S with k rows and n columns that satisfies the (ϵ, δ, ℓ)-JL moment property for some $\ell \geq 2$. Then for A, B matrices with n rows,

$$\Pr_S\left[\left|A^T S^T S B \ - A^T B\right|_F \geq 3\ \epsilon |A|_F |B|_F\right] \leq \delta$$

- Proof: For a random scalar X, let $|X|_p = (E|X|^p)^{1/p}$
  - Sometimes consider $X = |T|_F$ for a random matrix T
  - $|\ |T|_F\ |_p = \left(E\left[|T|_F^p\right]\right)^{1/p}$

- Can show $|.|_p$ is a norm if $p \geq 1$
  - Minkowski's Inequality: $|X + Y|_p \leq |X|_p + |Y|_p$

- For unit vectors x, y, we will bound $|\langle Sx, Sy\rangle$ - $\langle x, y\rangle|_\ell$

Minkowski's Inequality

- Minkowski's Inequality: $|X+Y|_p \leq |X|_p + |Y|_p$
- Proof:
  - If $|X|_p, |Y|_p$ are finite, then so is $|X+Y|_p$. Why?
  - $f(x) = x^p$ is convex for $p \geq 1$, so for any fixed x, y:

    $|.5x + .5y|^p \leq |.5|x| \quad + \ .5|y||^p \leq .5|x|^p + .5|y|^p$, so

    $$|x+y|^p \leq 2^{p-1}(|x|^p + |y|^p)$$
  - $\text{So}, E\left[|X+Y|_p^p\right] \leq E[2^{p-1}(|X|_p^p + |Y|_p^p)]$
  - $$\begin{aligned} |X+Y|_p^p &= \int |x+y|^p \, d\mu \\ &= \int |x+y| \cdot |x+y|^{p-1} d\mu \\ &\leq \int (|x| + |y|)|x+y|^{p-1} \, d\mu \\ &= \int |x||x+y|^{p-1} \, d\mu + \int |y||x+y|^{p-1} d\mu \\ &\leq \left(\left(\int |x|^p d\mu\right)^{\frac{1}{p}} + \left(\int |y|^p d\mu\right)^{\frac{1}{p}} \right) \left(\int |x+y|^{(p-1)\left(\frac{p}{p-1}\right)} d\mu \right)^{\frac{p-1}{p}} \\ &= \left(|X|_p + |Y|_p\right)|X+Y|_p^{p-1} \end{aligned}$$

From Vectors to Matrices

- For unit vectors x, y, $|\langle Sx, Sy\rangle - \langle x, y\rangle|_\ell$

$$= \frac{1}{2}\left|(|Sx|_2^2 - 1) + (|Sy|_2^2 - 1) - (|S(x-y)|_2^2 - |x-y|_2^2)\right|_\ell$$

$$\leq \frac{1}{2}\left(\left||Sx|_2^2 - 1\right|_\ell + \left||Sy|_2^2 - 1\right|_\ell + \left||S(x-y)|_2^2 - |x-y|_2^2\right|_\ell\right)$$

$$\leq \frac{1}{2}\left(\epsilon \cdot \delta^{\frac{1}{\ell}} + \epsilon \cdot \delta^{\frac{1}{\ell}} + |x-y|_2^2\, \epsilon \cdot \delta^{\frac{1}{\ell}}\right)$$

$$\leq 3\,\epsilon \cdot \delta^{\frac{1}{\ell}}$$

- By linearity, for arbitrary x, y, $\frac{|\langle Sx, Sy\rangle - \langle x, y\rangle|_\ell}{|x|_2 |y|_2} \leq 3\,\epsilon \cdot \delta^{\frac{1}{\ell}}$

- Suppose A has d columns and B has e columns. Let the columns of A be $A_1, \ldots, A_d$ and the columns of B be $B_1, \ldots, B_e$

- Define $X_{i,j} = \frac{1}{|A_i|_2 |B_j|_2} \cdot (\langle SA_i, SB_j\rangle - \langle A_i, B_j\rangle)$

- $\left|A^T S^T S B - A^T B\right|_F^2 = \sum_i \sum_j |A_i|_2^2 \cdot \left|B_j\right|_2^2 X_{i,j}^2$

From Vectors to Matrices

- Have shown: for arbitrary x, y, $\frac{|\langle Sx, Sy\rangle - \langle x,y\rangle|_\ell}{|x|_2 |y|_2} \leq 3\,\epsilon \cdot \delta^{\frac{1}{\ell}}$
- For $X_{i,j} = \frac{1}{|A_i|_2 |B_j|_2} \cdot \left(\langle SA_i, SB_j\rangle - \langle A_i, B_j\rangle\right)$: $\left|A^T S^T SB - A^T B\right|_F^2 = \sum_i \sum_j |A_i|_2^2 \cdot \left|B_j\right|_2^2 X_{i,j}^2$
- $\left|\left|A^T S^T SB - A^T B\right|_F^2\right|_{\ell/2} = \left|\sum_i \sum_j |A_i|_2^2 \cdot \left|B_j\right|_2^2 X_{i,j}^2\right|_{\ell/2}$

$$\leq \sum_i \sum_j |A_i|_2^2 \cdot \left|B_j\right|_2^2 |X_{i,j}^2|_{\ell/2}$$

$$= \sum_i \sum_j |A_i|_2^2 \cdot \left|B_j\right|_2^2 \left|X_{i,j}\right|_\ell^2$$

$$\leq \left(3\epsilon\delta^{\frac{1}{\ell}}\right)^2 \sum_i \sum_j |A_i|_2^2 \left|B_j\right|_2^2$$

$$= \left(3\,\epsilon\delta^{\frac{1}{\ell}}\right)^2 |A|_F^2 |B|_F^2$$

- Since $E\left[\left|A^T S^T SB - A^T B\right|_F^\ell\right] = \left|\left|A^T S^T SB - A^T B\right|_F^2\right|_{\frac{\ell}{2}}^{\ell/2}$, by Markov's inequality,
- $\Pr\left[\left|A^T S^T SB - A^T B\right|_F > 3\epsilon |A|_F |B|_F\right] \leq \left(\frac{1}{3\epsilon |A|_F |B|_F}\right)^\ell E\left[|A^T S^T SB - A^T B|_F^\ell\right] \leq \delta$

Result for Vectors

- Show: $\Pr[|CS^TSD - CD|_F^2 \leq [6/(\delta(\#\text{ rows of } S))] * |C|_F^2 |D|_F^2] \geq 1 - \delta$

- (JL Property) A distribution on matrices $S \in R^{k\times n}$ has the (ϵ, δ, ℓ)-JL moment property if for all $x \in R^n$ with $|x|_2 = 1$,

$$E_S \left| |Sx|_2^2 - 1 \right|^{\ell} \leq \epsilon^{\ell} \cdot \delta$$

- (From vectors to matrices) For $\epsilon, \delta \in \left(0, \frac{1}{2}\right)$, let D be a distribution on matrices S with k rows and n columns that satisfies the (ϵ, δ, ℓ)-JL moment property for some $\ell \geq 2$. Then for A, B matrices with n rows,

$$\Pr_S \left[\left| A^T S^T S B - A^T B \right|_F \geq 3\, \epsilon |A|_F |B|_F \right] \leq \delta$$

- Just need to show that the CountSketch matrix S satisfies JL property and bound the number k of rows

CountSketch Satisfies the JL Property

- (JL Property) A distribution on matrices $S \in R^{k x n}$ has the (ϵ, δ, ℓ)-JL moment property if for all $x \in R^n$ with $|x|_2 = 1$,

$$E_S \left| |Sx|_2^2 - 1 \right|^{\ell} \leq \epsilon^{\ell} \cdot \delta$$

- We show this property holds with $\ell = 2$. First, let us consider $\ell = 1$

- For CountSketch matrix S, let
  - h:[n] -> [k] be a 2-wise independent hash function
  - σ: $[n] \rightarrow \{-1,1\}$ be a 4-wise independent hash function

- Let $\delta(E) = 1$ if event E holds, and $\delta(E) = 0$ otherwise

- $E[|Sx|_2^2] = \sum_{j \in [k]} E[\left(\sum_{i \in [n]} \delta(h(i) = j)\sigma_i x_i\right)^2]$

$$= \sum_{j \in [k]} \sum_{i1, i2 \in [n]} E[\delta(h(i1) = j)\delta(h(i2) = j)\sigma_{i1}\sigma_{i2}] x_{i1} x_{i2}$$

$$= \sum_{j \in [k]} \sum_{i \in [n]} E[\delta(h(i) = j)^2] x_i^2$$

$$= \left(\frac{1}{k}\right) \sum_{j \in [k]} \sum_{i \in [n]} x_i^2 = |x|_2^2$$

CountSketch Satisfies the JL Property

- $E[|Sx|_2^4] = E[\sum_{j \in [k]} \sum_{j' \in [k]} \left(\sum_{i \in [n]} \delta(h(i) = j)\sigma_i x_i\right)^2 \left(\sum_{i' \in [n]} \delta(h(i') = j')\sigma_{i'} x_{i'}\right)^2] =$

  $\sum_{j_1, j_2, i_1, i_2, i_3, i_4} E[\sigma_{i1}\sigma_{i2}\sigma_{i3}\sigma_{i4}\delta(h(i_1) = j_1)\delta(h(i_2) = j_1)\delta(h(i_3) = j_2)\delta\big(h(i_4 = j_2)\big)] x_{i1} x_{i2} x_{i3} x_{i4}$

- We must be able to partition $\{i_1, i_2, i_3, i_4\}$ into equal pairs

- Suppose $i_1 = i_2 = i_3 = i_4$. Then necessarily $j_1 = j_2$. Obtain $\sum_j \frac{1}{k} \sum_i x_i^4 = |x|_4^4$

- Suppose $i_1 = i_2$ and $i_3 = i_4$ but $i_1 \neq i_3$. Then get $\sum_{j_1, j_2, i_1, i_3} \frac{1}{k^2} x_{i_1}^2 x_{i_3}^2 = |x|_2^4 - |x|_4^4$

- Suppose $i_1 = i_3$ and $i_2 = i_4$ but $i_1 \neq i_2$. Then necessarily $j_1 = j_2$. Obtain $\sum_j \frac{1}{k^2} \sum_{i_1, i_2} x_{i_1}^2 x_{i_2}^2 \leq \frac{1}{k} |x|_2^4$. Obtain same bound if $i_1 = i_4$ and $i_2 = i_3$.

- Hence, $E[|Sx|_2^4] \in [|x|_2^4, |x|_2^4(1 + \frac{2}{k})] = [1, 1 + \frac{2}{k}]$

- So, $E_S\left||Sx|_2^2 - 1\right|^2 \leq \left(1 + \frac{2}{k}\right) - 2 + 1 = \frac{2}{k}$. Setting $k = \frac{2}{\epsilon^2 \delta}$ finishes the proof

Where are we?

- (JL Property) A distribution on matrices $S \in R^{k x\, n}$ has the (ϵ, δ, ℓ)-JL moment property if for all $x \in R^n$ with $|x|_2 = 1$,
$$E_S \left| |Sx|_2^2 - 1 \right|^\ell \leq \epsilon^\ell \cdot \delta$$

- (From vectors to matrices) For $\epsilon, \delta \in \left(0, \frac{1}{2}\right)$, let D be a distribution on matrices S with k rows and n columns that satisfies the (ϵ, δ, ℓ)-JL moment property for some $\ell \geq 2$. Then for A, B matrices with n rows,
$$\Pr_S \left[\left| A^T S^T SB \; - A^T B \right|^2{}_F \geq 3\, \epsilon^2 \, |A|_F^2 |B|_F^2 \right] \leq \delta$$

- We showed CountSketch has the JL property with $\ell = 2, \text{ and } k = \frac{2}{\epsilon^2 \delta}$

- Matrix product result we wanted was:
  $\Pr[|CS^T SD - CD|_F^2 \leq (6/(\delta k)) * |C|_F^2 \, |D|_F^2] \geq 1 - \delta$
- We are now done with the proof CountSketch is a subspace embedding

Course Outline

- Subspace embeddings and least squares regression
  - Gaussian matrices
  - Subsampled Randomized Hadamard Transform
  - CountSketch
- Affine embeddings
  - Application to low rank approximation
- High precision regression
- Leverage score sampling
- Distributed low rank approximation
- L1 Regression
- M-Estimator regression

Affine Embeddings

- Want to solve $\min_X |AX - B|_F^2$, A is tall and thin with d columns, but B has a large number of columns
- Can't directly apply subspace embeddings
- Let's try to show $|SAX - SB|_F = (1 \pm \epsilon)|AX - B|_F$ for all X and see what properties we need of S
- Can assume A has orthonormal columns
- Let $B^* = AX^* - B$, where X^* is the optimum
- $|S(AX - B)|_F^2 - |SB^*|_F^2 = |SA(X - X^*) + S(AX^* - B)|_F^2 - |SB^*|_F^2$
  $= |SA(X - X^*)|_F^2 + 2\mathrm{tr}[(X - X^*)^T A^T S^T SB^*]$ (use $|C + D|_F^2 = |C|_F^2 + |D|_F^2 + 2\mathrm{Tr}(C^T D)$)
  $\in |SA(X - X^*)|_F^2 \pm 2|X - X^*|_F \left|A^T S^T SB^*\right|_F$ (use $\mathrm{tr}(CD) \le |C|_F |D|_F$)
  $\in |SA(X - X^*)|_F^2 \pm 2\epsilon |X - X^*|_F |B^*|_F$ (if we have approx. matrix product)
  $\in |A(X - X^*)|_F^2 \pm \epsilon(|A(X - X^*)|_F^2 + 2|X - X^*|_F |B^*|)$ (subspace embedding for A)

Affine Embeddings

- We have

$$|S(AX-B)|_F^2 - |SB^*|_F^2 \in |A(X-X^*)|_F^2 \pm \epsilon(|A(X-X^*)|_F^2 + 2|X-X^*|_F|B^*|)$$

- Normal equations imply that

$$|AX-B|_F^2 = |A(X-X^*)|_F^2 + |B^*|_F^2$$

- $|S(AX-B)|_F^2 - |SB^*|_F^2 - \left(|AX-B|_F^2 - |B^*|_F^2\right)$

$$\in \epsilon(|A(X-X^*)|_F^2 + 2|X-X^*|_F|B^*|_F)$$

$$\in \pm\epsilon\left(|A(X-X^*)|_F + |B^*|_F\right)^2$$

$$\in \pm 2\epsilon\left(|A(X-X^*)|_F{}^2 + |B^*|_F^2\right)$$

$$= \pm 2\epsilon|AX-B|_F^2$$

- $|SB^*|_F^2 = (1 \pm \epsilon)|B^*|_F^2$ (this holds with constant probability)

Affine Embeddings

- Know: $|S(AX-B)|_F^2 - |SB^*|_F^2 - \left(|AX-B|_F^2 - |B^*|_F^2\right) \in \pm 2\epsilon |AX-B|_F^2$
- Know: $|SB^*|_F^2 = (1 \pm \epsilon)|B^*|_F^2$
- $|S(AX-B)|_F^2 = (1 \pm 2\epsilon)|AX-B|_F^2 + \epsilon |B^*|_F^2$
  $= (1 \pm 3\epsilon)|AX-B|_F^2$
- Completes proof of affine embedding!

Affine Embeddings: Missing Proofs

- Claim: $|A + B|_F^2 = |A|_F^2 + |B|_F^2 + 2\mathrm{Tr}(A^T B)$

- Proof: $|A + B|_F^2 = \sum_i |A_i + B_i|_2^2$

$$= \sum_i |A_i|_2^2 + \sum_i |B_i|_2^2 + 2\langle A_i, B_i\rangle$$

$$= |A|_F^2 + |B|_F^2 + 2\mathrm{Tr}(A^T B)$$

Affine Embeddings: Missing Proofs

- Claim: $\mathrm{Tr}(AB) \leq |A|_F |B|_F$
- Proof: $\mathrm{Tr}(AB) = \sum_i \langle A^i, B_i \rangle$ for rows A^i and columns B_i

$\leq \sum_i \left|A^i\right|_2 |B_i|_2$ by Cauchy-Schwarz for each i

$\leq \left(\sum_i \left|A^i\right|_2^2\right)^{\frac{1}{2}} \left(\sum_i |B_i|_2^2\right)^{\frac{1}{2}}$ another Cauchy-Schwarz

$= |A|_F |B|_F$

Affine Embeddings: Homework Proof

- Claim: $|SB^*|_F^2 = (1 \pm \epsilon)|B^*|_F^2$ with constant probability if CountSketch matrix S has $k = O(\frac{1}{\epsilon^2})$ rows

- Proof:
- $|SB^*|_F^2 = \sum_i |SB_i^*|_2^2$

- By our analysis for CountSketch and linearity of expectation, $E\left[|SB^*|_F^2\right] = \sum_i E[|SB_i^*|_2^2] = |B^*|_F^2$

- $E[|SB^*|_F^4] = \sum_{i,j} E[|SB_i^*|_2^2 \left|SB_j^*\right|_2^2]$

- By our CountSketch analysis, $E[|SB_i^*|_2^4]] \leq |B_i^*|_2^4(1 + \frac{2}{k})$
- For cross terms see Lemma 40 in [CW13]

Low rank approximation

- A is an n x d matrix
  - Think of n points in R^d
- E.g., A is a customer-product matrix
  - $A_{i,j}$ = how many times customer i purchased item j
- A is typically well-approximated by low rank matrix
  - E.g., high rank because of noise
- Goal: find a low rank matrix approximating A
  - Easy to store, data more interpretable

What is a good low rank approximation?

Singular Value Decomposition (SVD)

Any matrix $A = U \cdot \Sigma \cdot V$

- U has orthonormal columns
- Σ is diagonal with non-increasing positive entries down the diagonal
- V has orthonormal rows

- Rank-k approximation: $A_k = U_k \cdot \Sigma_k \cdot V_k$
  - rows of V_k are the top k principal components

$$\begin{pmatrix} \\ \mathbf{A} \\ \\ \end{pmatrix} = \begin{pmatrix} \\ \mathbf{U}_k \\ \\ \end{pmatrix} \begin{pmatrix} \Sigma_k \end{pmatrix} \begin{pmatrix} \quad \mathbf{V}_k \quad \end{pmatrix} + \begin{pmatrix} \\ \quad \mathbf{E} \quad \\ \\ \end{pmatrix}$$

What is a good low rank approximation?

$A_k = \text{argmin}_{\text{rank k matrices B}} |A-B|_F$

$(|C|_F = (\Sigma_{i,j} C_{i,j}^2)^{1/2})$

Computing A_k exactly is expensive

$$\begin{pmatrix} & & \\ & \mathbf{A} & \\ & & \end{pmatrix} = \begin{pmatrix} \\ \mathbf{U}_k \\ \\ \end{pmatrix} \begin{pmatrix} \Sigma_k \end{pmatrix} \begin{pmatrix} & \mathbf{V}_k & \end{pmatrix} + \begin{pmatrix} & & \\ & \mathbf{E} & \\ & & \end{pmatrix}$$

Low rank approximation

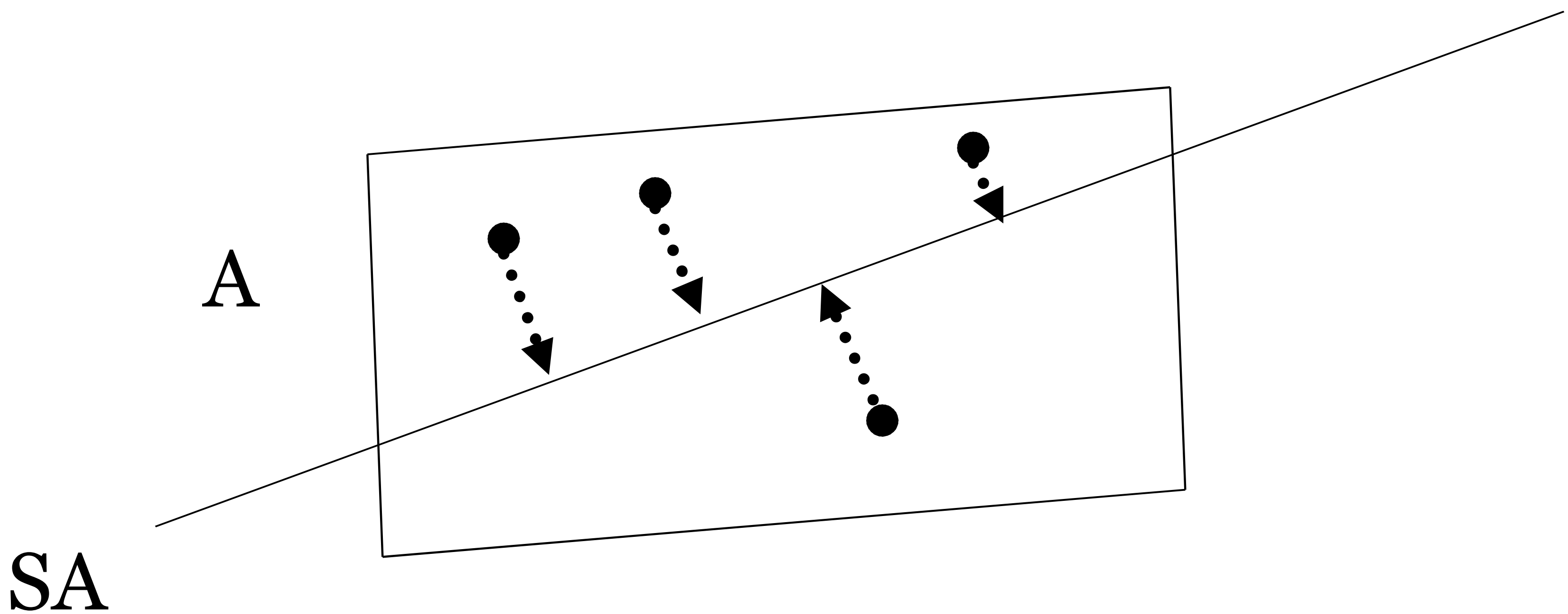
- Goal: output a rank k matrix A', so that

$$|A-A'|_F \leq (1+\varepsilon)\,|A-A_k|_F$$

- Can do this in nnz(A) + (n+d)*poly(k/ε) time [S,CW]
  - nnz(A) is number of non-zero entries of A

Solution to low-rank approximation [S]

- Given n x d input matrix A
- Compute S*A using a random matrix S with k/ε << n rows. S*A takes random linear combinations of rows of A



- Project rows of A onto SA, then find best rank-k approximation to points inside of SA.

What is the matrix S?

- S can be a k/ε x n matrix of i.i.d. normal random variables
- [S] S can be a k/ε x n Fast Johnson Lindenstrauss Matrix
  - Uses Fast Fourier Transform
- [CW] S can be a poly(k/ε) x n CountSketch matrix

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$S \cdot A$ can be computed in nnz(A) time

Why do these Matrices Work?

- Consider the regression problem $\min_X |A_k X - A|_F$
- Let S be an affine embedding
- Then $|SA_k X - SA|_F = (1 \pm \epsilon)|A_k X - A|_F$ for all X
- By normal equations, $\operatorname{argmin}_X |SA_k X - SA|_F = (SA_k)^- SA$
- So, $|A_k (SA_k)^- SA - A|_F \leq (1 + \epsilon)|A_k - A|_F$
- But $A_k (SA_k)^- SA$ is a rank-k matrix in the row span of SA!
- Let’s formalize why the algorithm works now…

Why do these Matrices Work?

- $\min_{\text{rank}-k\ X} |XSA - A|_F^2 \leq \left|A_k(SA_k)^- SA - A\right|_F^2 \leq (1+\epsilon)\left|A - A\_k\right|_F^2$

- By the normal equations,
$$|XSA - A|_F^2 = |XSA - A(SA)^- SA|_F^2 + |A(SA)^- SA - A|_F^2$$

- Hence,
$$\min_{\text{rank}-k\ X} |XSA - A|_F^2 = |A(SA)^- SA - A|_F^2 + \min_{\text{rank}-k\ X} |XSA - A(SA)^- SA|_F^2$$

- Can write $SA = U\,\Sigma V^T$ in its SVD

- Then, $\min_{\text{rank}-k\ X} |XSA - A(SA)^- SA|_F^2 = \min_{\text{rank}-k\ X} |XU\Sigma - A(SA)^- U\Sigma|_F^2$
$$= \min_{\text{rank}-k\ Y} |Y - A(SA)^- U\Sigma|_F^2$$

- Hence, we can just compute the SVD of $A(SA)^- U\Sigma$

- But how do we compute $A(SA)^- U\Sigma$ quickly?

Caveat: projecting the points onto SA is slow

- Current algorithm:
  1. Compute S*A
  2. Project each of the rows onto S*A
  3. Find best rank-k approximation of projected points inside of rowspace of S*A
- Bottleneck is step 2

$\min_{\text{rank-k } X} |X(SA)R-AR|_F^2$

Can solve with affine embeddings

- [CW] Approximate the projection
  - Fast algorithm for approximate regression

    $\min_{\text{rank-k } X} |X(SA)-A|_F^2$

  - Want nnz(A) + (n+d)*poly(k/ε) time

Using Affine Embeddings

- We know we can just output $\arg \min_{\text{rank}-k\ X} |XSA - A|_F^2$

- Choose an affine embedding R:

$$|XSAR - AR|_F^2 = (1 \pm \epsilon)|XSA - A|_F^2 \text{ for all X}$$

- Note: we can compute AR and SAR in nnz(A) time

- Can just solve $\min_{\text{rank}-k\ X} |XSAR - AR|_F^2$

- $\min_{\text{rank}-k\ X} |XSAR - AR|_F^2 = |AR(SAR)^{-}(SAR) - AR|_F^2 + \min_{\text{rank}-k\ X} |XSAR - AR(SAR)^{-}(SAR)|_F^2$

- Compute $\min_{\text{rank}-k\ Y} |Y - AR(SAR)^{-}(SAR)|_F^2$ using SVD which is $(n + d)\text{poly}\left(\frac{k}{\epsilon}\right)$ time

- Necessarily, $Y = XSAR$ for some X. Output $Y(SAR)^{-}SA$ in factored form. We're done!

Low Rank Approximation Summary

1. Compute SA

2. Compute SAR and AR

3. Compute $\min_{\text{rank}-k\, Y} |Y - AR(SAR)^{-}(SAR)|_F^2$ using SVD

4. Output $Y(SAR)^{-}SA$ in factored form

Overall time: nnz(A) + (n+d)poly(k/ε)

Course Outline

- Subspace embeddings and least squares regression
  - Gaussian matrices
  - Subsampled Randomized Hadamard Transform
  - CountSketch
- Affine embeddings
  - Application to low rank approximation
- High precision regression
- Leverage score sampling
- Distributed low rank approximation
- L1 Regression
- M-Estimator regression

High Precision Regression

- Goal: output x' for which $|Ax'-b|_2 \leq (1+\varepsilon) \min_x |Ax-b|_2$ with high probability
- Our algorithms all have running time poly(d/ε)
- Goal: Sometimes we want running time poly(d)*log(1/ε)
- Want to make A well-conditioned
  - $\kappa(A) = \sup_{|x|_2=1} |Ax|_2 / \inf_{|x|_2=1} |Ax|_2$
- Lots of algorithms' time complexity depends on $\kappa(A)$
- Use sketching to reduce $\kappa(A)$ to O(1)!

Small QR Decomposition

- Let S be a $(1+\epsilon_0)$- subspace embedding for A
- Compute SA
- Compute QR-factorization, $\mathrm{SA} = \mathrm{QR}^{-1}$
- Claim: $\kappa(\mathrm{AR}) = \frac{(1+\epsilon_0)}{1-\epsilon_0}$
- For all unit x, $(1-\epsilon_0)|\mathrm{ARx}|_2 \leq |\mathrm{SAR\,x}|_2 = 1$
- For all unit x, $(1+\epsilon_0)|\mathrm{ARx}|_2 \geq |\mathrm{SARx}|_2 = 1$
- So $\kappa(\mathrm{AR}) = \sup_{|x|_2=1} |\mathrm{ARx}|_2 / \inf_{|x|_2=1} |\mathrm{ARx}|_2 \leq \frac{1+\epsilon_0}{1-\epsilon_0}$

Finding a Constant Factor Solution

- Let S be a $1+\epsilon_0$ - subspace embedding for AR
- Solve $x_0 = \underset{x}{\operatorname{argmin}} |SARx - Sb|_2$
- Time to compute R and x_0 is nnz(A) + poly(d) for constant ϵ_0
- $x_{m+1} \leftarrow x_m + R^T A^T (b - AR\, x_m)$
- $$\begin{aligned} AR(x_{m+1} - x^*) &= AR\,(x_m + R^T A^T (b - ARx_m) - x^*) \\ &= \left(AR - ARR^T A^T AR\right)(x_m - x^*) \\ &= U(\Sigma - \Sigma^3) V^T (x_m - x^*), \end{aligned}$$
  where $AR = U\,\Sigma V^T$ is the SVD of AR
- $|AR(x_{m+1} - x^*)|_2 = \left|(\Sigma - \Sigma^3)V^T\left(x_m - x^*\right)\right|_2 = O(\epsilon_0)|AR(x_m - x^*)|_2$
- $|ARx_m - b|^2{}_2 = |AR(x_m - x^*)|_2^2 + |ARx^* - b|_2^2$

Course Outline

- Subspace embeddings and least squares regression
  - Gaussian matrices
  - Subsampled Randomized Hadamard Transform
  - CountSketch
- Affine embeddings
  - Application to low rank approximation
- High precision regression
- Leverage score sampling
- Distributed low rank approximation
- M-Estimator regression

Leverage Score Sampling

- This is another subspace embedding, but it is based on sampling!
  - If A has sparse rows, then SA has sparse rows!

- Let $A = U\,\Sigma V^T$ be an n x d matrix with rank d, written in its SVD

- Define the i-th leverage score $\ell(i)$ of A to be $|U_{i,*}|_2^2$

- What is $\sum_i \ell(i)$?
  - Let $(q_1, \ldots, q_n)$ be a distribution with $q_i \geq \frac{\beta \ell(i)}{d}$, where β is a parameter

- Define sampling matrix $S = D \cdot \Omega^T$, where D is k x k and Ω is n x k
  - Ω is a sampling matrix, and D is a rescaling matrix
  - For each column j of Ω, D, independently, and with replacement, pick a row index i in [n] with probability q_i, and set $\Omega_{i,j} = 1$ and $D_{j,j} = 1/(q_i k)^{.5}$

Leverage Score Sampling

- Note: leverage scores do not depend on choice of orthonormal basis U for columns of A
- Indeed, let U and U' be two such orthonormal bases
- Claim: $|\mathrm{e_i U}|_2^2 = |\mathrm{e_i U'}|_2^2$ for all i
- Proof: Since both U and U' have column space equal to that of A, we have $\mathrm{U = U'Z}$ for change of basis matrix Z
- Since U and U' each have orthonormal columns, Z is a rotation matrix (orthonormal rows and columns)
- Then $|\mathrm{e_i U}|_2^2 = |\mathrm{e_i U'Z}|_2^2 = |\mathrm{e_i U'}|_2^2$

Leverage Score Sampling gives a Subspace Embedding

- Want to show for $S = D \cdot \Omega^T$, that $|SAx|_2^2 = (1 \pm \epsilon)|Ax|_2^2$ for all x
- Writing $A = U\Sigma V^T$ in its SVD, this is equivalent to showing $|SUy|_2^2 = (1 \pm \epsilon)|Uy|_2^2 = (1 \pm \epsilon)|y|_2^2$ for all y
- As usual, we can just show with high probability, $|U^T S^T SU - I|_2 \le \epsilon$
- How can we analyze $U^T S^T SU$?
- (Matrix Chernoff) Let $X_1, \ldots, X_k$ be independent copies of a symmetric random matrix $X \in R^{dxd}$ with $E[X] = 0$, $|X|_2 \le \gamma$, and $|E[X^T X]|_2 \le \sigma^2$. Let $W = \frac{1}{k}\sum_{j \in [k]} X_j$. For any $\epsilon > 0$,

$$\Pr[|W|_2 > \epsilon] \le 2d \cdot e^{-k\epsilon^2/(\sigma^2 + \frac{\gamma\epsilon}{3})}$$

(here $|W|_2 = \sup \frac{|Wx|_2}{|x|_2}$. Since W is symmetric, $|W|_2 = \sup_{|x|_2 = 1} x^T W x$.)

Leverage Score Sampling gives a Subspace Embedding

- Let i(j) denote the index of the row of U sampled in the j-th trial
- Let $X_j = I_d - \frac{U_{i(j)}^T U_{i(j)}}{q_{i(j)}}$, where $U_{i(j)}$ is the j-th sampled row of U
- The X_j are independent copies of a symmetric matrix random variable
- $E[X_j] = I_d - \sum_i q_i \left(\frac{U_i^T U_i}{q_i}\right) = I_d - I_d = 0^d$
- $|X_j|_2 \le |I_d|_2 + \frac{|U_{i(j)}^T U_{i(j)}|_2}{q_{i(j)}} \le 1 + \max_i \frac{|U_i|_2^2}{q_i} \le 1 + \frac{d}{\beta}$
- $E[X^T X] = I_d - 2E\left[\frac{U_{i(j)}^T U_{i(j)}}{q_{i(j)}}\right] + E\left[\frac{U_{i(j)}^T U_{i(j)} U_{i(j)}^T U_{i(j)}}{q_{i(j)}^2}\right]$

$$= \sum_i \frac{U_i^T U_i U_i^T U_i}{q(i)} - I_d \le \left(\frac{d}{\beta}\right) \sum_i U_i^T U_i - I_d \le \left(\frac{d}{\beta} - 1\right) I_d,$$

where $A \le B$ means $x^T A x \le x^T B x$ for all x

- Hence, $|E[X^T X]|_2 \le \frac{d}{\beta} - 1$

Applying the Matrix Chernoff Bound

- (Matrix Chernoff) Let $X_1, \ldots, X_k$ be independent copies of a symmetric random matrix $X \in R^{dxd}$ with $E[X] = 0$, $|X|_2 \leq \gamma$, and $\left|E[X^T X]\right|_2 \leq \sigma^2$. Let $W = \frac{1}{k}\sum_{j \in [k]} X_j$. For any $\epsilon > 0$,

$$\Pr[|W|_2 > \epsilon] \leq 2d \cdot e^{-k\epsilon^2/(\sigma^2 + \frac{\gamma\epsilon}{3})}$$

(here $|W|_2 = \sup \frac{|Wx|_2}{|x|_2}$. Since W is symmetric, $|W|_2 = \sup_{|x|_2 = 1} x^T W x$.)

- $\gamma = 1 + \frac{d}{\beta}$, and $\sigma^2 = \frac{d}{\beta} - 1$

- $X_j = I_d - \frac{U_{i(j)}^T U_{i(j)}}{q_{i(j)}}$, and recall how we generated $S = D \cdot \Omega^T$: For each column j of Ω, D, independently, and with replacement, pick a row index i in [n] with probability q_i, and set $\Omega_{i,j} = 1$ and $D_{j,j} = 1/(q_i k)^{.5}$
  - Implies $W = I_d - U^T S^T S U$

- $\Pr\left[\left|I_d - U^T S^T S U\right|_2 > \epsilon\right] \leq 2d \cdot e^{-k\epsilon^2 \Theta\left(\frac{\beta}{d}\right)}$. Set $k = \Theta(\frac{d \log d}{\beta\epsilon^2})$ and we're done.

Fast Computation of Leverage Scores

- Naively, need to do an SVD to compute leverage scores
- Suppose we compute SA for a subspace embedding S
- Let $SA = QR^{-1}$ be such that Q has orthonormal columns
- Set $\ell_i' = |e_i AR|_2^2$
- Since AR has the same column span of A, $AR = UT^{-1}$
  - $(1-\epsilon)|ARx|_2 \leq |SARx|_2 = |x|_2$
  - $(1+\epsilon)|ARx|_2 \geq |SARx|_2 = |x|_2$
  - $(1 \pm O(\epsilon))|x|_2 = |ARx|_2 = |UT^{-1}x|_2 = |T^{-1}x|_2,$
- $\ell_i \ = |e_i ART|_2^2 = (1 \pm O(\epsilon))|e_i AR|_2^2 = (1 \pm O(\epsilon))\ell_i{}'$
- But how do we compute AR? We want nnz(A) time

Fast Computation of Leverage Scores

- $\ell_i = (1 \pm O(\epsilon))\ell_i{}'$
- Suffices to set this ϵ to be a constant
- Set $\ell_i' = |e_i AR|_2^2$
  - This takes too long
- Let G be a d x O(log n) matrix of i.i.d. normal random variables
  - For any vector z, $\Pr[|zG|_2^2 = \left(1 \pm \frac{1}{2}\right)|z|^2] \geq 1 - \frac{1}{n^2}$
- Instead set $\ell_i' = |e_i ARG|_2^2$.
  - Can compute in (nnz(A) + d^2) $\log n$ time
- Can solve regression in nnz(A) log n + poly(d(log n)/ε) time

Course Outline

- Subspace embeddings and least squares regression
  - Gaussian matrices
  - Subsampled Randomized Hadamard Transform
  - CountSketch
- Affine embeddings
  - Application to low rank approximation
- High precision regression
- Leverage score sampling
- Distributed low rank approximation
- L1 Regression
- M-Estimator regression

Distributed low rank approximation

- *We have fast algorithms for low rank approximation, but can they be made to work in a distributed setting?*

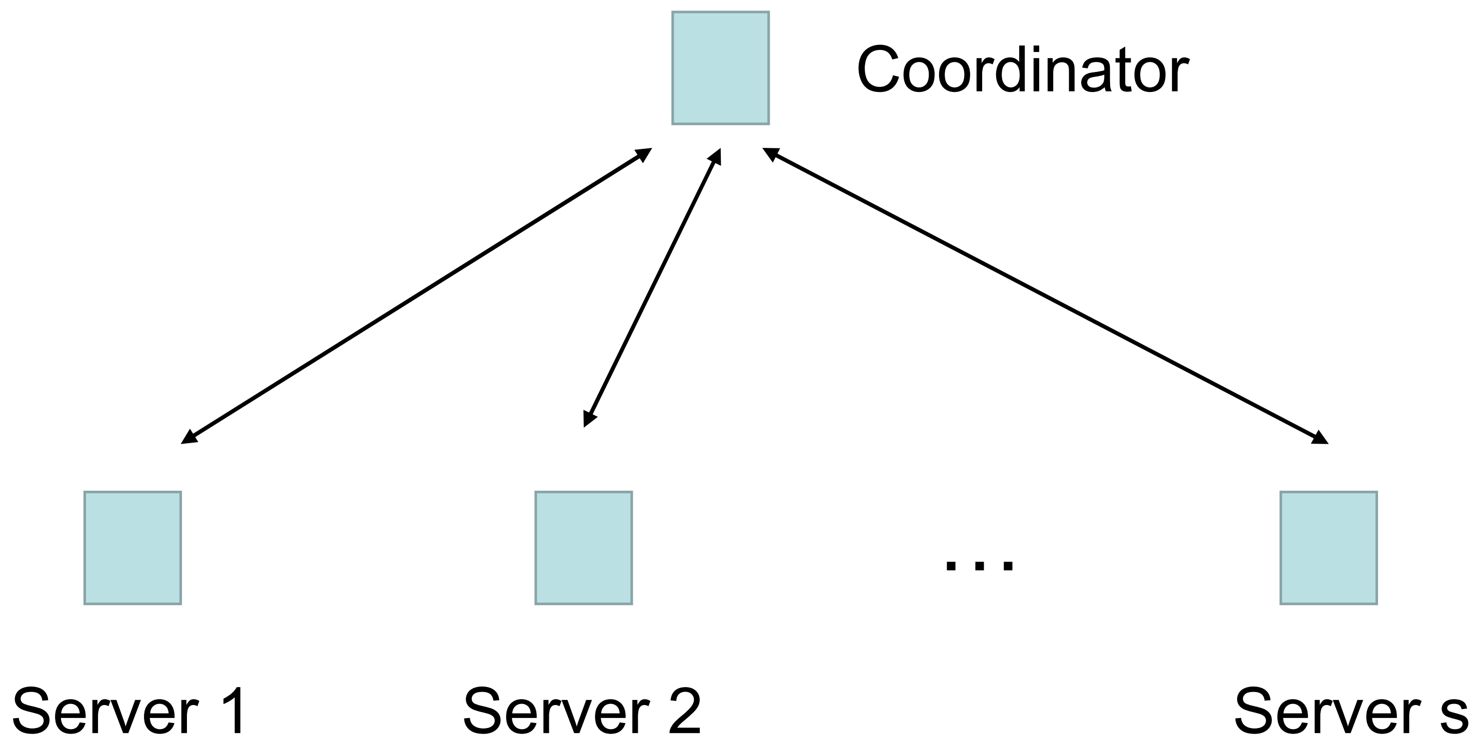
- Matrix A distributed among s servers

- For t = 1, …, s, we get a customer-product matrix from the t-th shop stored in server t. Server t's matrix = A^t

- Customer-product matrix $A = A^1 + A^2 + \dots + A^s$
  - Model is called the arbitrary partition model

- More general than the row-partition model in which each customer shops in only one shop

The Communication Model



- Each player talks only to a Coordinator via 2-way communication
- Can simulate arbitrary point-to-point communication up to factor of 2 (and an additive O(log s) factor per message)

Communication cost of low rank approximation

- Input: n x d matrix A stored on s servers
  - Server t has n x d matrix A^t
  - $A = A^1 + A^2 + \ldots + A^s$
  - Assume entries of A^t are O(log(nd))-bit integers
- Output: Each server outputs the same k-dimensional space W
  - $C = A^1 P_W + A^2 P_W + \ldots + A^s P_W$, where P_W is the projection onto W
  - $|A-C|_F \leq (1+\varepsilon)|A-A_k|_F$
  - Application: k-means clustering
- Resources: Minimize total communication and computation. Also want O(1) rounds and input sparsity time

Work on Distributed Low Rank Approximation

- [FSS]: First protocol for the row-partition model.
  - $O(sdk/\varepsilon)$ real numbers of communication
  - Don't analyze bit complexity (can be large)
  - SVD Running time, see also [BKLW]

- [KVW]: $O(skd/\varepsilon)$ communication in arbitrary partition model

- [BWZ]: $O(skd) + \mathrm{poly}(sk/\varepsilon)$ words of communication in arbitrary partition model. Input sparsity time
  - Matching $\Omega(skd)$ words of communication lower bound

- Variants: kernel low rank approximation [BLSWX], low rank approximation of an implicit matrix [WZ], sparsity [BWZ]

Outline of Distributed Protocols

- [FSS] protocol
- [KVW] protocol
- [BWZ] protocol

Constructing a Coreset [FSS]

- Let $A = U\,\Sigma V^T$ be its SVD
- Let m = k + k/ϵ
- Let Σ_m agree with Σ on the first m diagonal entries, and be 0 otherwise
- Claim: For all projection matrices Y=I-X onto (d-k)-dimensional subspaces,

$$\left|\Sigma_m V^T Y\right|_F^2 = (1 \pm \epsilon)|AY|_F^2 + c,$$

where $c = |A - A_m|_F^2$ does not depend on Y

- We can think of S as U_m^T so that $SA = U_m^T U\Sigma V^T = \Sigma_m V^T$ is a sketch

Constructing a Coreset

- Claim: For all projection matrices Y=I-X onto (n-k)-dimensional subspaces,

$$\left|\Sigma_m V^T Y\right|_F^2 + c \ = (1 \pm \epsilon)|AY|_F^2,$$

where $c = |A - A_m|_F^2$ does not depend on Y

- Proof: $|AY|_F^2 = \left|U\Sigma_m V^T Y\right|_F^2 + \left|U(\Sigma - \Sigma_m)V^T Y\right|_F^2$

$$\leq \left|\Sigma_m V^T Y\right|_F^2 + |A - A_m|_F^2 = \left|\Sigma_m V^T Y\right|_F^2 + c$$

Also, $\left|\Sigma_m V^T Y\right|_F^2 + |A - A_m|_F^2 \ - |AY|_F^2$

$$= \left|\Sigma_m V^T\right|_F^2 - \left|\Sigma_m V^T X\right|_F^2 + |A - A_m|_F^2 - |A|_F^2 + |AX|_F^2$$

$$= |AX|_F^2 - \left|\Sigma_m V^T X\right|_F^2$$

$$= \left|(\Sigma - \Sigma_m)V^T X\right|_F^2$$

$$\leq \left|(\Sigma - \Sigma_m)V^T\right|_2^2 \cdot |X|_F^2$$

$$\leq \sigma_{m+1}^2 \, k \leq \epsilon \, \sigma_{m+1}^2 (m - k) \leq \epsilon \sum_{i \in \{k+1,..,m+1\}} \sigma_i^2 \leq \epsilon |A - A_k|_F^2$$

Unions of Coresets

- Suppose we have matrices $A^1, \ldots, A^s$ and construct $\Sigma_m^1 V^{T,1}, \Sigma_m^2 V^{T,2}, \ldots, \Sigma_m^s V^{T,s}$ as in the previous slide, together with $c_1, \ldots, c_s$

- Then $\sum_i \quad \left|\Sigma_m^i V^{T,i} Y\right|_F^2 + c_i \ = (1 \pm \epsilon)|AY|_F^2$, where A is the matrix formed by concatenating the rows of $A^1, \ldots, A^s$

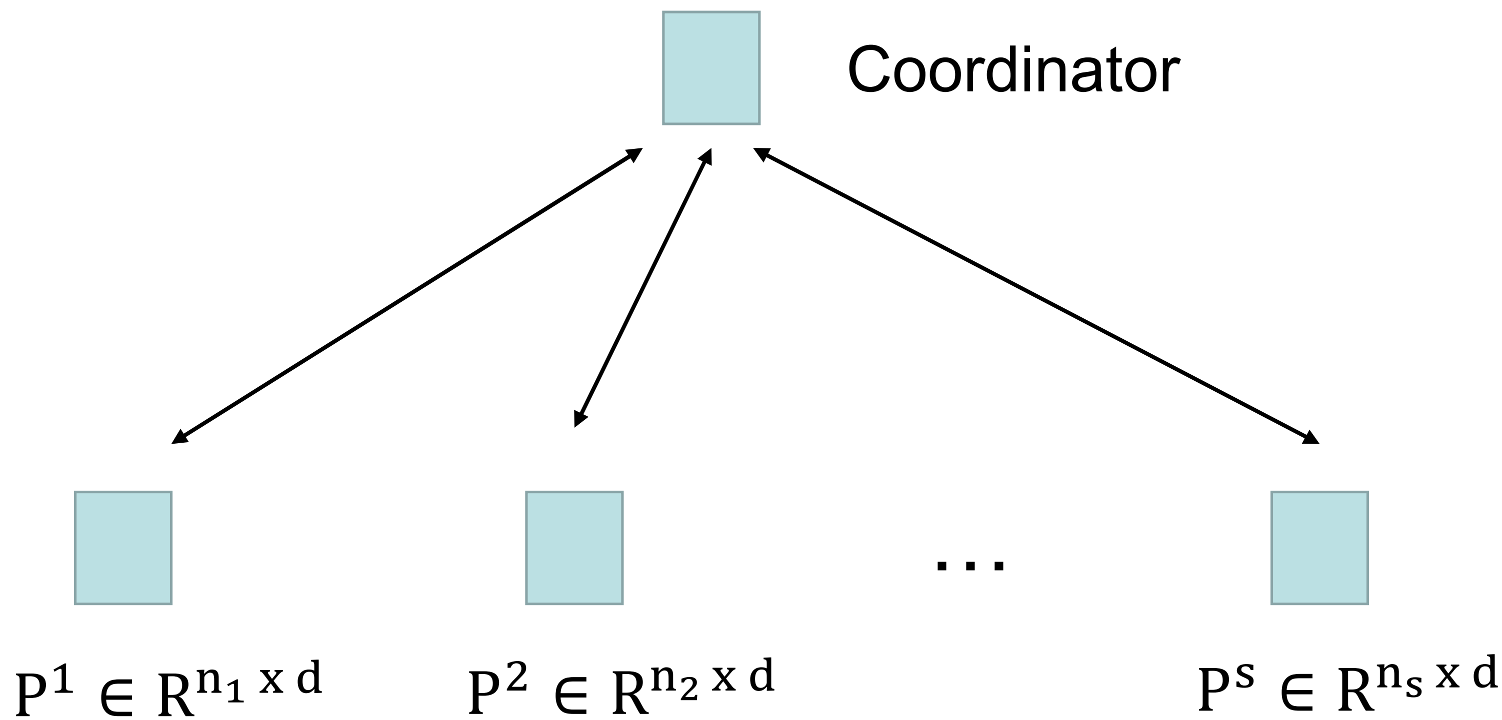
- Let B be the matrix obtained by concatenating the rows of $\Sigma_m^1 V^{T,1}, \Sigma_m^2 V^{T,2}, \ldots, \Sigma_m^s V^{T,s}$

- Suppose we compute $B = U\,\Sigma V^T$ and compute $\Sigma_m V^T$ and $|B - B_m|_F^2$

- Then $\left|\Sigma_m V^T Y\right|_F^2 + c + \sum_i c_i = (1 \pm \epsilon)|BY|_F^2 + \sum_i c_i = (1 \pm O(\epsilon))|AY|_F^2$

- So $\Sigma_m V^T$ and the constant $c + \sum_i c_i$ are a coreset for A

[FSS] Row-Partition Protocol



- Server t sends the top k/ε + k principal components of P^t, scaled by the top k/ε + k singular values Σ^t, together with c^t
- Coordinator returns top k principal components of $[\Sigma^1 V^1; \Sigma^2 V^2; \ldots; \Sigma^s V^s]$

[FSS] Row-Partition Protocol

[KVW] protocol will handle 2, 3, and 4

Problems:

1. sdk/ε real numbers of communication
2. bit complexity can be large
3. running time for SVDs [BLKW]
4. doesn't work in arbitrary partition model

This is an SVD-based protocol. Maybe our random matrix techniques can improve communication just like they improved computation?

[KVW] Arbitrary Partition Model Protocol

- Inspired by the sketching algorithm presented earlier
- Let S be one of the k/ε x n random matrices discussed
  - S can be generated pseudorandomly from small seed
  - Coordinator sends small seed for S to all servers
- Server t computes SA^t and sends it to Coordinator
- Coordinator sends $\Sigma_{t=1}^{s} SA^t = SA$ to all servers
- There is a good k-dimensional subspace inside of SA. If we knew it, t-th server could output projection of A^t onto it

[KVW] Arbitrary Partition Model Protocol

Problems:

- Can't output projection of A^t onto SA since the rank is too large

- Could communicate this projection to the coordinator who could find a k-dimensional space, but communication depends on n

[KVW] Arbitrary Partition Model Protocol

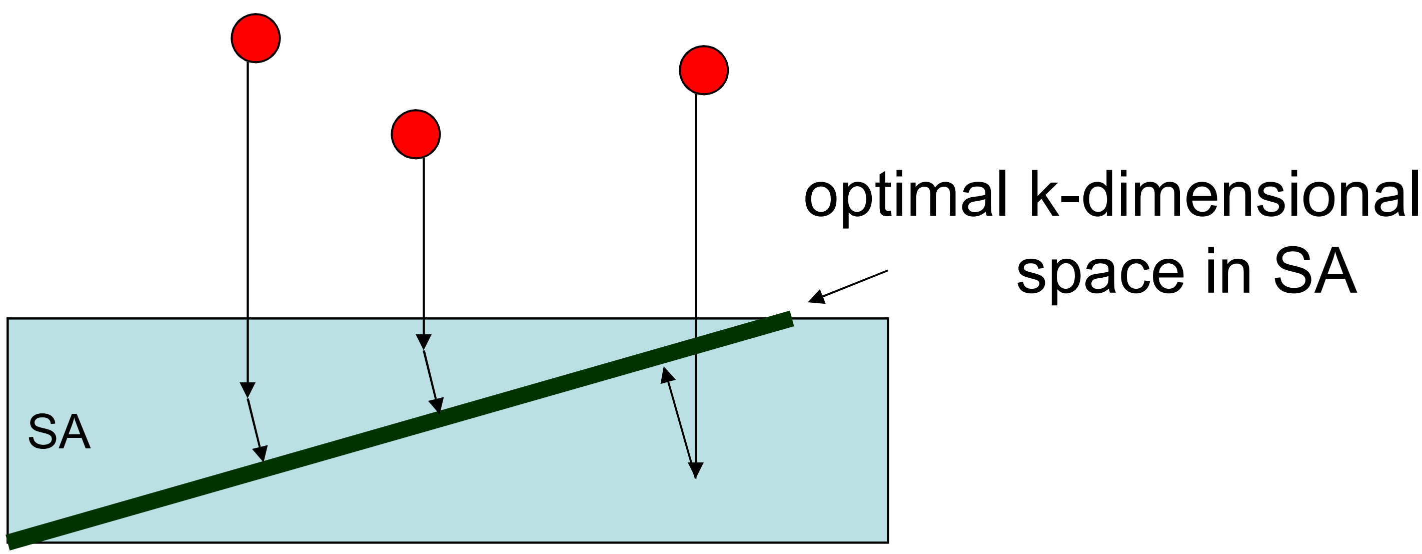
Fix:

- Instead of projecting A onto SA, recall we can solve $\min_{\text{rank-k } X} \left|A(SA)^T XSA - A\right|_F^2$
- Let T_1, T_2 be affine embeddings, solve $\min_{\text{rank-k } X} \left|T_1 A(SA)^T XSAT_2 - T_1 AT_2\right|_F^2$ (optimization problem is small and has a closed form solution)
- Everyone can then compute XSA and then output k directions

[KVW] protocol

- Phase 1:

- Learn the row space of SA

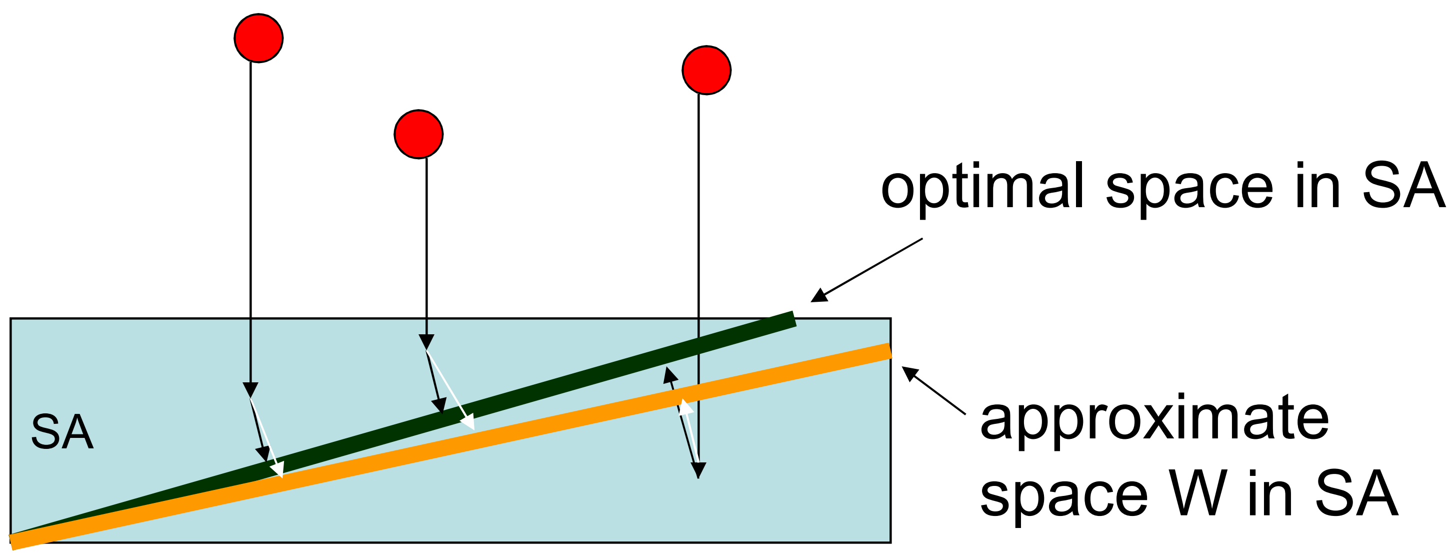


cost $\leq (1+\varepsilon)|A-A_k|_F$

[KVW] protocol

- Phase 2:

- Find an approximately optimal space W inside of SA



cost $\leq (1+\varepsilon)^2|A-A_k|_F$

[BWZ] Protocol

- Main Problem: communication is O(skd/ε) + poly(sk/ε)

- We want O(skd) + poly(sk/ε) communication!

- Idea: use projection-cost preserving sketches [CEMMP]

- Let A be an n x d matrix

- If S is a random k/ε^2 x n matrix, then there is a constant $c \geq 0$ so that for all k-dimensional projection matrices P:

$$|\mathrm{SA}(\mathrm{I}-\mathrm{P})|_{\mathrm{F}} + \mathrm{c} = (1 \pm \epsilon)|\mathrm{A}(\mathrm{I}-\mathrm{P})|_{\mathrm{F}}$$

[BWZ] Protocol

Intuitively, U looks like top k left singular vectors of SA

- Let S be a k/ε^2 x n projection-cost preserving sketch
- Let T be a d x k/ε^2 projection-cost preserving sketch
- Server t sends SA^tT to Coordinator
- Coordinator sends back SAT = $\sum_t SA^tT$ to servers
- Each server computes k/ε^2x k matrix U of top k left singular vectors of SAT

Thus, U^TSA looks like top k right singular vectors of SA

- Server t sends U^TSA^t to Coordinator
- Coordinator returns the space $U^TSA = \sum_t U^TSA^t$ to output

Top k right singular vectors of SA work because S is a projection-cost preserving sketch!

[BWZ] Analysis

- Let W be the row span of $U^T SA$, and P be the projection onto W
- Want to show $|A - AP|_F \le (1+\epsilon)|A - A_k|_F$
- Since T is a projection-cost preserving sketch,

$$(*) \quad |SA - SAP|_F \le \left|SA - UU^T SA\right|_F + c_1 \le (1+\epsilon)|SA - [SA]_k|_F$$

- Since S is a projection-cost preserving sketch, there is a scalar c > 0, so that for all k-dimensional projection matrices Q,

$$|SA - SAQ|_F + c = (1 \pm \epsilon)|A - AQ|_F$$

- Add c to both sides of (*) to conclude $|A - AP|_F \le (1+\epsilon)|A - A_k|_F$

Conclusions for Distributed Low Rank Approximation

- [BWZ] Optimal O(sdk) + poly(sk/ε) communication protocol for low rank approximation in arbitrary partition model
  - Handle bit complexity by adding noise
  - Input sparsity time
  - 2 rounds, which is optimal [W]
  - Optimal data stream algorithms improves [CW, L, GP]

- Communication of other optimization problems?
  - Computing the rank of an n x n matrix over the reals
  - Linear Programming
  - Graph problems: Matching
  - etc.

Course Outline

- Subspace embeddings and least squares regression
  - Gaussian matrices
  - Subsampled Randomized Hadamard Transform
  - CountSketch
- Affine embeddings
  - Application to low rank approximation
- High precision regression
- Leverage score sampling
- Distributed low rank approximation
- L1 Regression
- M-Estimator Regression

Robust Regression

Method of least absolute deviation (l_1 -regression)

- Find x* that minimizes $|Ax-b|_1 = \Sigma\ |b_i - \langle A_{i*}, x\rangle|$
- Cost is less sensitive to outliers than least squares
- Can solve via linear programming

Solving l_1 -regression via Linear Programming

- Minimize $(1,\ldots,1) \cdot (\alpha^+ + \alpha^-)$
- Subject to:

$$A x + \alpha^+ - \alpha^- = b$$
$$\alpha^+, \alpha^- \geq 0$$

- Generic linear programming gives poly(nd) time
- Want much faster time using sketching!

Well-Conditioned Bases

- For an n x d matrix A, can choose an n x d matrix U with orthonormal columns for which A = UW, and $|\mathrm{Ux}|_2 = |\mathrm{x}|_2$ for all x

- Can we find a U for which A = UW and $|\mathrm{Ux}|_1 \approx |\mathrm{x}|_1$ for all x?

- Let A = QW where Q has full column rank, and define $|\mathrm{z}|_{\mathrm{Q},1} = |\mathrm{Qz}|_1$
  - $|\mathrm{z}|_{\mathrm{Q},1}$ is a norm

- Let C = $\{\mathrm{z} \in \mathrm{R}^{\mathrm{d}} : |\mathrm{z}|_{\mathrm{Q},1} \leq 1\}$ be the unit ball of $|.|_{\mathrm{Q},1}$

- C is a convex set which is symmetric about the origin
  - Lowner-John Theorem: can find an ellipsoid E such that: $\mathrm{E} \subseteq \mathrm{C} \subseteq \sqrt{\mathrm{d}}\mathrm{E}$, where E = $\{\mathrm{z} \in \mathrm{R}^{\mathrm{d}} : \mathrm{z}^{\mathrm{T}}\mathrm{Fz} \leq 1\}$
  - $(\mathrm{z}^{\mathrm{T}}\mathrm{Fz})^{.5} \leq |\mathrm{z}|_{\mathrm{Q},1} \leq \sqrt{\mathrm{d}}(\mathrm{z}^{\mathrm{T}}\mathrm{Fz})^{.5}$
  - $\mathrm{F} = \mathrm{GG}^{\mathrm{T}}$ since F defines an ellipsoid

- Define $\mathrm{U} = \mathrm{QG}^{-1}$

Well-Conditioned Bases

- Recall $\mathrm{U} = \mathrm{QG}^{-1}$ where $(\mathrm{z}^{\mathrm{T}}\mathrm{Fz})^{.5} \leq |\mathrm{z}|_{\mathrm{Q},1} \leq \sqrt{\mathrm{d}}(\mathrm{z}^{\mathrm{T}}\mathrm{Fz})^{.5}$ and $\mathrm{F} = \mathrm{GG}^{\mathrm{T}}$
- $|\mathrm{Ux}|_1 = |\mathrm{QG}^{-1}\mathrm{x}|_1 = |\mathrm{Qz}|_1 = |\mathrm{z}|_{\mathrm{Q},1}$ where $\mathrm{z} = \mathrm{G}^{-1}\mathrm{x}$
- $\mathrm{z}^{\mathrm{T}}\mathrm{Fz} \quad = \left(\mathrm{x}^{\mathrm{T}}(\mathrm{G}^{-1})^{\mathrm{T}}\mathrm{G}^{\mathrm{T}}\mathrm{G}\,(\mathrm{G}^{-1})\mathrm{x}\right) = \mathrm{x}^{\mathrm{T}}\mathrm{x} = |\mathrm{x}|_2^2$
- So $|\mathrm{x}|_2 \leq |\mathrm{Ux}|_1 \leq \sqrt{\mathrm{d}}|\mathrm{x}|_2$
- So $\frac{|\mathrm{x}|_1}{\sqrt{\mathrm{d}}} \leq |\mathrm{x}|_2 \leq |\mathrm{Ux}|_1 \leq \sqrt{\mathrm{d}}|\mathrm{x}|_2 \leq \sqrt{\mathrm{d}}|\mathrm{x}|_1$

Net for ℓ_1 – Ball

- Consider the unit ℓ_1-ball B = $\{x \in R^d : |x|_1 = 1\}$
- Subset N is a γ-net if for all $x \in B$, there is a $y \in N$, such that $|x - y|_1 \leq \gamma$
- Greedy construction of N
  - While there is a point $x \in B$ of distance larger than γ from every point in N, include x in N
- The ℓ_1-ball of radius γ/2 around every point in N is contained in the ℓ_1-ball of radius 1+ γ/2 around 0^d
- Further, all such ball are disjoint
- Ratio of volume of d-dimensional similar polytopes of radius 1+ γ/2 to radius $\gamma/2$ is $(1 + \gamma/2)^d/(\gamma/2)^d$, so $|N| \leq (1 + \gamma/2)^d/(\gamma/2)^d$

Net for $\ell_1 - \text{Subspace}$

- Let A = UW for a well-conditioned basis U
  - $|x|_1 \leq |Ux|_1 \leq d|x|_1$ for all x
- Let N be a (γ/d) $-$net for the unit ℓ_1-ball B
- Let M = {Ux | x in N}, so $|M| \leq (1+\gamma/(2d))^d/(\gamma/(2d))^d$
- Claim: For every x in B, there is a y in M for which $|Ax - y|_1 \leq \gamma$
- Proof: Let x' in B be such that $|x - x'|_1 \leq \gamma/d$

  Then $|Ax - Ax'|_1 \leq d|x - x'|_1 \leq \gamma$, using the well-conditioned basis property. Set y = Ax'
- $|M| \leq \left(\frac{d}{\gamma}\right)^{O(d)}$

Rough Algorithm Overview

Compute poly(d)-approximation

Compute well-conditioned basis

Sample rows from the well-conditioned basis and the residual of the poly(d)-approximation

Takes nnz(A) time

Takes poly(d/ε) time

Solve l_1-regression on the sample, obtaining vector x, and output x

Rough Algorithm Overview

$\min_{x \text{ in } R^d} |Ax-b|_1 = \min_{x \text{ in } R^d} |Ux - b'|_1$

Sample poly(d/ε) rows of $U \circ b'$ proportional to their l_1-norm.

Find x' such that
$|Ax'-b|_1 \leq \text{poly}(d) \min_{x \text{ in } R^d} |Ax-b|_1$
Let b' = b-Ax' be the residual

Find a basis A=UW so that for all x in R^d,
$|x|_1/\text{poly}(d) \leq |Ux|_1 \leq \text{poly}(d)\ |x|_1$

Now generic linear programming is efficient

Will focus on showing how to quickly compute

1. A poly(d)-approximation

2. A well-conditioned basis

Sketching Theorem

Theorem

- There is a probability space over $(d \log d) \times n$ matrices R such that for any $n \times d$ matrix A, with probability at least 99/100 we have for all x:

$$|Ax|_1 \leq |RAx|_1 \leq d \log d \cdot |Ax|_1$$

Embedding

- is linear
- is independent of A
- preserves lengths of an infinite number of vectors

Application of Sketching Theorem

Computing a d(log d)-approximation

- Compute RA and Rb
- Solve $x' = \text{argmin}_x |RAx-Rb|_1$
- Main theorem applied to $A \circ b$ implies x' is a d log d – approximation
- RA, Rb have d log d rows, so can solve l_1-regression efficiently

Application of Sketching Theorem

Computing a well-conditioned basis

1. Compute RA
2. Compute W so that RAW is orthonormal (in the l_2-sense)
3. Output U = AW

U = AW is well-conditioned because

$|AWx|_1 \leq |RAWx|_1 \leq (d \log d)^{1/2} |RAWx|_2 = (d \log d)^{1/2} |x|_2 \leq (d \log d)^{1/2} |x|_1$

and

$|AWx|_1 \geq |RAWx|_1/(d \log d) \geq |RAWx|_2/(d \log d) = |x|_2/(d \log d) \geq |x|_1 /(d^{3/2} \log d)$

Sketching Theorem

Theorem:

- There is a probability space over $(d \log d) \times n$ matrices R such that for any $n \times d$ matrix A, with probability at least 99/100 we have for all x:

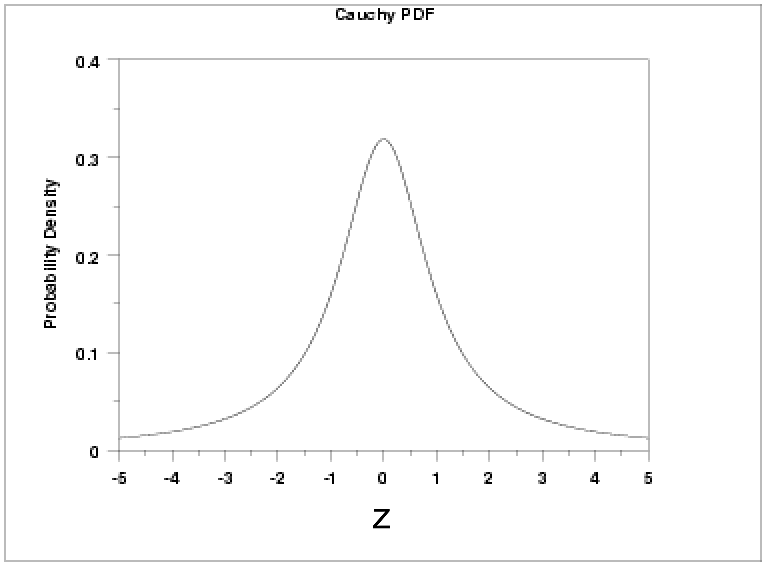
$$|Ax|_1 \leq |RAx|_1 \leq d \log d \cdot |Ax|_1$$

A dense R that works:

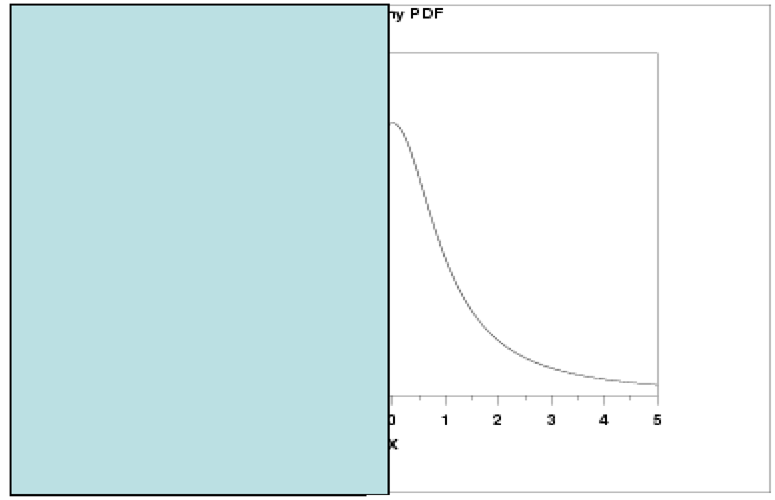
The entries of R are i.i.d. Cauchy random variables, scaled by $1/(d \log d)$

Cauchy Random Variables

- pdf(z) = $1/(\pi(1+z^2))$ for z in $(-\infty, \infty)$
- Undefined expectation and infinite variance
- 1-stable:
  - If $z_1, z_2, \ldots, z_n$ are i.i.d. Cauchy, then for $a \in R^n$,
    $a_1 \cdot z_1 + a_2 \cdot z_2 + \ldots + a_n \cdot z_n \sim |a|_1 \cdot z$, where z is Cauchy
- Can generate as the ratio of two standard normal random variables



Proof of Sketching Theorem



z

- By 1-stability,
  - For all rows r of R,
    - $\langle r, Ax \rangle = |Ax|_1 \cdot Z / (d \log d)$,

      where Z is a Cauchy
- $RAx = (|Ax|_1 \cdot Z_1, \ldots, |Ax|_1 \cdot Z_{d \log d}) / (d \log d)$,

  where $Z_1, \ldots, Z_{d \log d}$ are i.i.d. Cauchy
- $|RAx|_1 = |Ax|_1 \sum_j |Z_j| / (d \log d)$
  - The $|Z_j|$ are half-Cauchy
- $\sum_j |Z_j| = \Omega(d \log d)$ with probability $1-\exp(-d \log d)$ by Chernoff
- But the $|Z_j|$ are heavy-tailed…

Proof of Sketching Theorem

- $\sum_j |Z_j|$ is heavy-tailed, so $|RAx|_1 = |Ax|_1 \sum_j |Z_j| / (d \log d)$ may be large
- Each $|Z_j|$ has c.d.f. asymptotic to $1-\Theta(1/z)$ for z in $[0, \infty)$
- There *exists* a well-conditioned basis of A
  - Suppose w.l.o.g. the basis vectors are $A_{*1}, \ldots, A_{*d}$
- $|RA_{*i}|_1 = |A_{*i}|_1 \cdot \sum_j |Z_{i,j}| / (d \log d)$
- Let $E_{i,j}$ be the event that $|Z_{i,j}| \leq d^3$
  - Define $Z'_{i,j} = |Z_{i,j}|$ if $|Z_{i,j}| \leq d^3$, and $Z'_{i,j} = d^3$ otherwise
  - $E\left[Z_{i,j} \mid E_{i,j}\right] = E\left[Z'_{i,j} \mid E_{i,j}\right] = O(\log d)$
- Let E be the event that for all i,j, $E_{i,j}$ occurs
  - $\Pr[E] \geq 1 - \frac{\log d}{d}$
- What is $E\left[Z'_{i,j} \mid E\right]$?

Proof of Sketching Theorem

- What is $\mathrm{E}[Z'_{i,j} \mid \mathrm{E}]$?

- $$\begin{aligned} \mathrm{E}[Z'_{i,j} | \mathrm{E}_{i,j}] &= \mathrm{E}[Z'_{i,j} | \mathrm{E}_{i,j}, \mathrm{E}] \Pr[\mathrm{E} \mid \mathrm{E}_{i,j}] + \mathrm{E}[Z'_{i,j} | \mathrm{E}_{i,j}, \neg\mathrm{E}] \Pr[\neg\mathrm{E} \mid \mathrm{E}_{i,j}] \\ &\geq \mathrm{E}[Z'_{i,j} | \mathrm{E}_{i,j}, \mathrm{E}] \Pr[\mathrm{E} \mid \mathrm{E}_{i,j}] \\ &= \mathrm{E}[Z'_{i,j} | \mathrm{E}] \cdot \left(\frac{\Pr[\mathrm{E}_{i,j} | \mathrm{E}] \Pr[\mathrm{E}]}{\Pr[\mathrm{E}_{i,j}]} \right) \\ &\geq \mathrm{E}[Z'_{i,j} | \mathrm{E}] \cdot \left(1 - \frac{\log d}{d} \right) \end{aligned}$$

- So, $\mathrm{E}[Z'_{i,j} | \mathrm{E}] = O(\log d)$
- $|RA_{*i}|_1 = |A_{*i}|_1 \cdot \sum_{i,j} |Z_{i,j}| / (d \log d)$
- With constant probability, $\sum_i |RA_{*i}|_1 = O(\log d) \sum_i |A_{*i}|_1$

Proof of Sketching Theorem

- With constant probability, $\sum_i |RA_{*i}|_1 = O(\log d) \sum_i |A_{*i}|_1$
- Recall $A_{*1}, \ldots, A_{*d}$ is a well-conditioned basis, and we showed the existence of such a basis earlier
- We will use the Auerbach basis which always exists:
  - For all x, $|x|_\infty \leq |Ax|_1$
  - $\sum_i |A_{*i}|_1 = d$
- $\sum_i |RA_{*i}|_1 = O(d \log d)$
- For all x, $|RAx|_1 \leq \sum_i |RA_{*i} x_i| \leq |x|_\infty \sum_i |RA_{*i}|_1$

  $= |x|_\infty O(d \log d)$

  $= O(d \log d) |Ax|_1$

Where are we?

- Suffices to show for all x with $|x|_1 = 1$, that $|Ax|_1 \leq |RAx|_1 \leq d \log d \cdot |Ax|_1$
- We know
  - (1) there is a γ-net M, with $|M| \leq \left(\frac{d}{\gamma}\right)^{O(d)}$, of the set {Ax such that $|x|_1 = 1$}
  - (2) for any fixed x, $|RAx|_1 \geq |Ax|_1$ with probability $1 - \exp(-d \log d)$
  - (3) for all x, $|RAx|_1 = O(d \log d)|Ax|_1$
- Set $\gamma = 1/(d^3 \log d)$ so $|M| \leq d^{O(d)}$
  - By a union bound, for all y in M, $|Ry|_1 \geq |y|_1$
- Let x with $|x|_1 = 1$ be arbitrary. Let y in M satisfy $|Ax - y|_1 \leq \gamma = 1/(d^3 \log d)$
- $|RAx|_1 \geq |Ry|_1 - |R(Ax - y)|_1$

$$\begin{aligned}
&\geq |y|_1 - O(d \log d)|Ax - y|_1 \\
&\geq |y|_1 - O(d \log d)\gamma \\
&\geq |y|_1 - O\left(\frac{1}{d^2}\right) \\
&\geq |y|_1/2 \quad \text{(why?)}
\end{aligned}$$

Sketching to solve l_1-regression [CW, MM]

- Most expensive operation is computing R*A where R is the matrix of i.i.d. Cauchy random variables

- All other operations are in the "smaller space"

- Can speed this up by choosing R as follows:

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} C_1 & & & & \\ & C_2 & & & \\ & & C_3 & & \\ & & & \ldots & \\ & & & & C_n \end{bmatrix}$$

Further sketching improvements [WZ]

- Can show you need a fewer number of sampled rows in later steps if instead choose R as follows

- Instead of diagonal of Cauchy random variables, choose diagonal of reciprocals of exponential random variables

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1/E_1 & & & & \\ & 1/E_2 & & & \\ & & 1/E_3 & & \\ & & & \cdots & \\ & & & & 1/E_n \end{bmatrix}$$

Course Outline

- Subspace embeddings and least squares regression
  - Gaussian matrices
  - Subsampled Randomized Hadamard Transform
  - CountSketch
- Affine embeddings
  - Application to low rank approximation
- High precision regression
- Leverage score sampling
- Distributed low rank approximation
- L1 Regression
- M-Estimator Regression

Robust Regression Fitness Measures

Example: Method of least absolute deviation (l_1 -regression)

- Find x* that minimizes $|Ax-b|_1 = \Sigma |b_i - <A_{i*}, x>|$
- Cost is less sensitive to outliers than least squares
- Can solve via linear programming
- Can solve in nnz(A) + poly(d/ε) time using sketching

What about the many other fitness measures used in practice?

M-Estimators

- Measure function
  - $M: R \rightarrow R^{0}$
  - $M(x) = M(-x)$, $M(0) = 0$
  - M is non-decreasing in $|x|$

- $|y|_M = \Sigma_{i=1}^{n} M(y_i)$

- Solve $\min_x |Ax-b|_M$

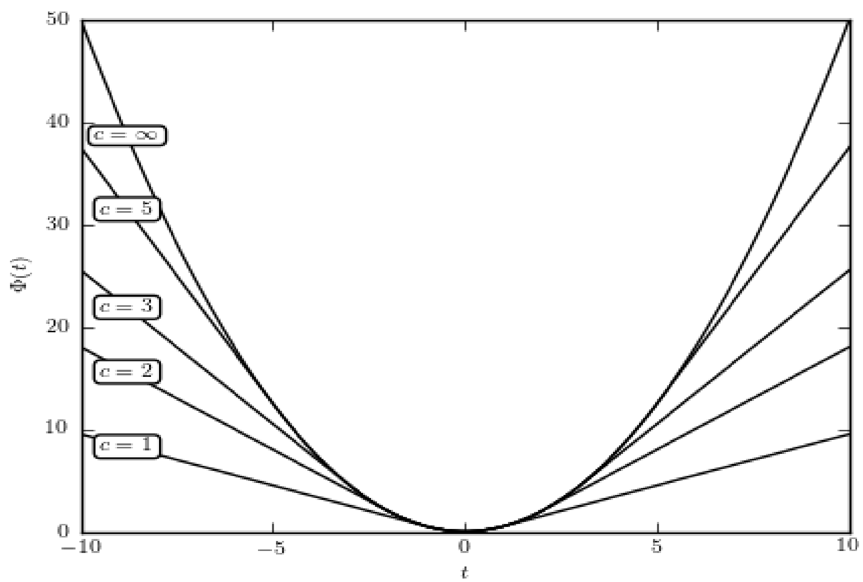
- Least squares and L_1-regression are special cases

Huber Loss Function

$M(x) = x^2/(2c)$ for $|x| \leq c$

$M(x) = |x|-c/2$ for $|x| > c$

Enjoys smoothness properties of l_2^2 and robustness properties of l_1



Other Examples

- L_1-L_2

  $$M(x) = 2((1+x^2/2)^{1/2} - 1)$$

- Fair estimator

  $$M(x) = c^2 [\ |x|/c\ - \log(1+|x|/c)\ ]$$

- Tukey estimator

  $$\begin{aligned} M(x) &= c^2/6\ (1-[1-(x/c)^2]^3) && \text{if } |x| \leq c \\ &= c^2/6 && \text{if } |x| > c \end{aligned}$$

Nice M-Estimators

- An M-Estimator is nice if it has at least linear growth and at most quadratic growth

- There is $C_M > 0$ so that for all a, a' with $|a| \geq |a'| > 0$,

$$|a/a'|^2 \geq M(a)/M(a') \geq C_M \, |a/a'|$$

- Any convex M satisfies the linear lower bound (why?)

$$M(a') = M\left(\left(\frac{a'}{a}\right) \cdot a + \left(1 - \frac{a'}{a}\right) \cdot 0\right) \leq \left(\frac{a'}{a}\right) M(a) + \left(1 - \frac{a'}{a}\right) M(0) = \left(\frac{a'}{a}\right) M(a)$$

- Any sketchable M satisfies the quadratic upper bound
  - sketchable => there is a distribution on k x n matrices S for which $|Sx|_M = \Theta(|x|_M)$ with good probability and k is slow-growing function of n

Nice M-Estimator Theorem

[Nice M-Estimators] O(nnz(A)) + poly(d log n) time algorithm to output x' so that for any constant C > 1, with probability 99%:

$$|Ax'-b|_M \leq C \min_x |Ax-b|_M$$

Remarks:

- For convex nice M-estimators can solve with convex programming, but slow – poly(nd) time

- Our sketch is "universal"

M-Sketch

$$T = \begin{bmatrix} S^0 \cdot D^0 \\ S^1 \cdot D^1 \\ S^2 \cdot D^2 \\ \cdots \\ S^{\log n} \cdot D^{\log n} \end{bmatrix}$$

- S^i are independent CountSketch matrices with poly(d) rows
- D^i is n x n diagonal and uniformly samples a $1/(d \log n)^i$ fraction of the n rows

-The same M-Sketch works for all nice M-estimators!

$$x' = \text{argmin}_x \; |TAx-Tb|_{w,M}$$

- many analyses of this data structure don't work since they reduce the problem to a non-convex problem

- Sketch used for estimating frequency moments [Indyk, W] and earthmover distance [Verbin, Zhang]

M-Sketch Intuition

- For a given y = Ax-b, consider $|Ty|_{w, M} = \Sigma_i w_i M((Ty)_i)$
- [Contraction] $|Ty|_{w,M} \geq ½ |y|_M$ with probability 1-exp(-d log n)
- [Dilation] $|Ty|_{w,M} \leq 2 |y|_M$ with probability 99%
- Contraction allows for a net argument (no scale-invariance!)
  - Show that $|y^*|_2$ is within a factor poly(n) of $\min_x |Ax - b|_2$
- Dilation implies the optimal y* does not dilate much
- Proof: try to estimate contribution to $|y|_M$ at all scales
  - E.g., if y = (n, 1, 1, …, 1) with a total of n-1 1s, then $|y|_1$ = n + (n-1)*1
  - When estimating a given scale, use the fact that smaller stuff cancels each other out in a bucket and gives its 2-norm